DOCTOR OF PHILOSOPHY

Computational Analysis of High-Replicate RNA-seq Data in Saccharomyces cerevisiae Searching for New Genomic Features

Copeland, Nancy Giang

Award date:
2018

COMPUTATIONAL ANALYSIS OF HIGH-REPLICATE RNA-SEQ DATA IN *SACCHAROMYCES CEREVISIAE*: SEARCHING FOR NEW GENOMIC FEATURES

By

Nancy Giang Copeland

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

AT

UNIVERSITY OF DUNDEE

DUNDEE, UNITED KINGDOM

MAY 2018

Contents

List of Tables

List of Figures

Acknowledgements

I would to first and foremost thank Geoff for giving me the opportunity to pursue a PhD in his research group. Thank you to both Nick and Pieta for providing me with such great support and being such phenomenal mentors in research and life.

I would like to give a special thanks to Nick and Rhoda for always being a listening ear and helping me keep my sanity through board games and the laughter of their happy children.

Thank you to Suzanne for so many encouraging words and for helping me through that final push. Our coffee breaks always ended too soon.

To Simon and Roger - you have always reassured me in times of doubt, continuously showing me that it is possible to accomplish this mortal task.

Thank you to Thiago and Fabio for being my amazing shoulders to lean on. You kept me going with your banter, cheerfulness, and those nifty new tips and tricks you always seem to find.

Tara - thank you for showing me the ropes and for helping me realise that I was doing just fine. I am truly grateful for your mentorship and friendship.

A big thanks go to Chris and Jim for your vast knowledge and sarcasm. Where would the group or any of us be without you guys.

Thank you to Joe and Leanne for taking such good care of me. I am truly grateful for your kindness and continual understanding.

To all of my friends and family in the UK - I want to thank each and every one of you for being such a wonderful network of support and encouragement.

And finally, I give my biggest thanks to my husband Tom. You have been by my side through this entire journey with your unwavering patience, support, and encouragement. I cannot thank you enough, and I really could not have done it without you.

To Thomas Martin Copeland

. . .

and Professor Chun Wai Liew.

UNIVERSITY OF DUNDEE

COLLEGE OF LIFE SCIENCES

I certify that Nancy Giang has satisfied all the terms and conditions of the relevant Ordinance and Regulations to qualify in submitting this thesis in application for the degree of Doctor of Philosophy.

Dated: May 2018

Research Supervisor: _____

Professor Geoffrey J. Barton

UNIVERSITY OF DUNDEE

Date: **May 2018**

Author: **Nancy Giang Copeland**

Title: **Computational Analysis of High-Replicate RNA-seq Data in *Saccharomyces cerevisiae*: Searching for New Genomic Features**

Department: **College of Life Sciences**

Degree: **Ph.D.**

I hereby declare that the work described in this thesis is my own; that I am the author of this thesis; that it has not previously been put forward in submission for any other degree or qualification; and that I have consulted references herein.

Nancy Giang Copeland

Abbreviations

aa = amino acid(s)

ACS = autonomously replicating sequence consensus sequence

ARS = autonomously replicating sequence

BLAST = Basic Local Alignment Search Tool

blastn = nucleotide query to search nucleotide databases with BLAST

blastp = protein query to search protein databases with BLAST

blastx = translated nucleotide query to search protein databases with BLAST

bp = base pair

cDNA = complementary DNA

ChIP = chromatin immunoprecipitation

ChIP-chip = ChIP with DNA microarray

ChIP-exo = ChIP with exonucleases

ChIP-seq = ChIP sequencing

CUT = cryptic unstable transcript

Da = Dalton

DNA = deoxyribonucleic acid

DRS = direct RNA-seq

FDR = false discovery rate

G1 = Growth/Gap 1 phase of cell cycle

G2 = pre-mitotic phase of cell cycle

HMM = Hidden Markov model

mL = milliliter

mRNA = messenger RNA

NCBI = National Center for Biotechnology Information

ncRNA = non-coding RNA

nm = nanometer

ORF = open reading frame

PARS = parallel analysis of RNA structure

polyA = polyadenylation

RNA = ribonucleic acid

RPKM = reads per kilobase per million reads mapped

RPM = reads per million reads mapped

rRNA = ribosomal RNA

SAGE = serial analysis of gene expression

SDS-PAGE= sodium dodecyl sulfate polyacrylamide gel electrophoresis

SGD = *Saccharomyces* Genome Database

SILAC = stable isotope labelling of amino acids in cell culture

sORF = small open reading frame

snoRNA = small nucleolar RNA

snRNA = small nuclear RNA

SUT = stable unannotated (or uncharacterised) transcript

tblastn = protein query to search translated nucleotide databases with BLAST

tblastx = translated nucleotide query to search translated nucleotide databases with BLAST

TSS = transcription start site

UAR = un-annotated region

UHPLC = ultra-high performance liquid chromatography

uORF = upstream open reading frame

UTR = un-translated region

WT = wild-type

XUT = Xrn1-sensitive unstable transcript

YPAD = yeast extract-peptone-dextrose medium + adenine

Abstract

In this study, RNA-seq and proteomics, two orthogonal high-throughput technologies, were used to search the *Saccharomyces cerevisiae* genome for new genomic features. RNA-seq data were aligned to the genome with three successively stringent set of parameters for the STAR aligner (Dobin et al., 2013). The varying levels of stringency elucidated some complexities in the RNA-seq data, such as the presence of read alignments that mapped to multiple genomic locations. The RNA-seq alignments indicated the presence of RNA transcripts derived from regions of the genome without annotations (un-annotated regions) in the *Saccharomyces* Genome Database (SGD). To ensure that all of the high-quality curated annotations within SGD were accounted for appropriately, these datasets were categorised as either Primary or Secondary Annotations. Annotations of genomic regions where the primary sequence produced a molecule (e.g. snoRNA or peptide) were designated as Primary. Annotations of regions where other types of activity were present (e.g. histone binding sites, double-strand break hotspots) were classified as Secondary. Only the Primary Annotations were used as boundaries for determining locations of un-annotated regions. Open reading frames (ORFs) were present in these un-annotated regions. Therefore, the regions were translated in six frames to build a

database of all theoretical peptides. Proteomics tandem mass spectra were then searched against this peptide database to find the presence of any expressed ORFs within the un-annotated regions. Two preliminary target ORFs have been found to contain RNA-seq alignments and were detected by the proteomics analysis, evidence that their transcripts may have been present in the original sample. The next step would be to verify these two preliminary target regions in the experimental laboratory to determine if they are in fact expressed as peptides, and if so, what possible functions the peptides may have. Throughout this study, the Un-Annotated Region Pipeline (UAR-Pipeline) software was constructed to facilitate the analysis of un-annotated regions given a genome sequence, a set of genomic annotations, and RNA-seq data. In addition, a Quickload Site within the Integrated Genome Browser (Nicol et al., 2009) was created to store and effectively visualise un-annotated regions against RNA-seq alignments, annotations, and other tracks of information such as conservation. The vast majority of annotations contained within the Quickload Site are also hosted by SGD; therefore, the Site would serve as a new resource for the research community through anticipated public access.

Chapter 1

Literature Review

1.1 Introduction

This study uses two orthogonal high-throughput technologies, RNA-sequencing and proteomics, to search for new genomic features in *Saccharomyces cerevisiae*. RNA-sequencing provides information regarding the transcriptome of the organism, whereas proteomics characterises the collection of peptides and proteins present. If an RNA transcript is seen in the RNA-seq data and its corresponding peptide is found in the proteomics analysis, there is more yet evidence that the transcript is expressed.

A concurrent objective was characterising nucleic acid and protein sequences of interest found throughout the course of the study with current bioinformatics tools. The use of programs such as BLAST (Altschul et al., 1990), phylogenetic trees (Arthur Lesk, 2008), Hidden Markov Models (Krogh et al., 1994), phastCons (Siepel et al., 2005), and InterPro (Apweiler et al., 2001) are introduced and described later in this chapter.

1.2 Central Dogma

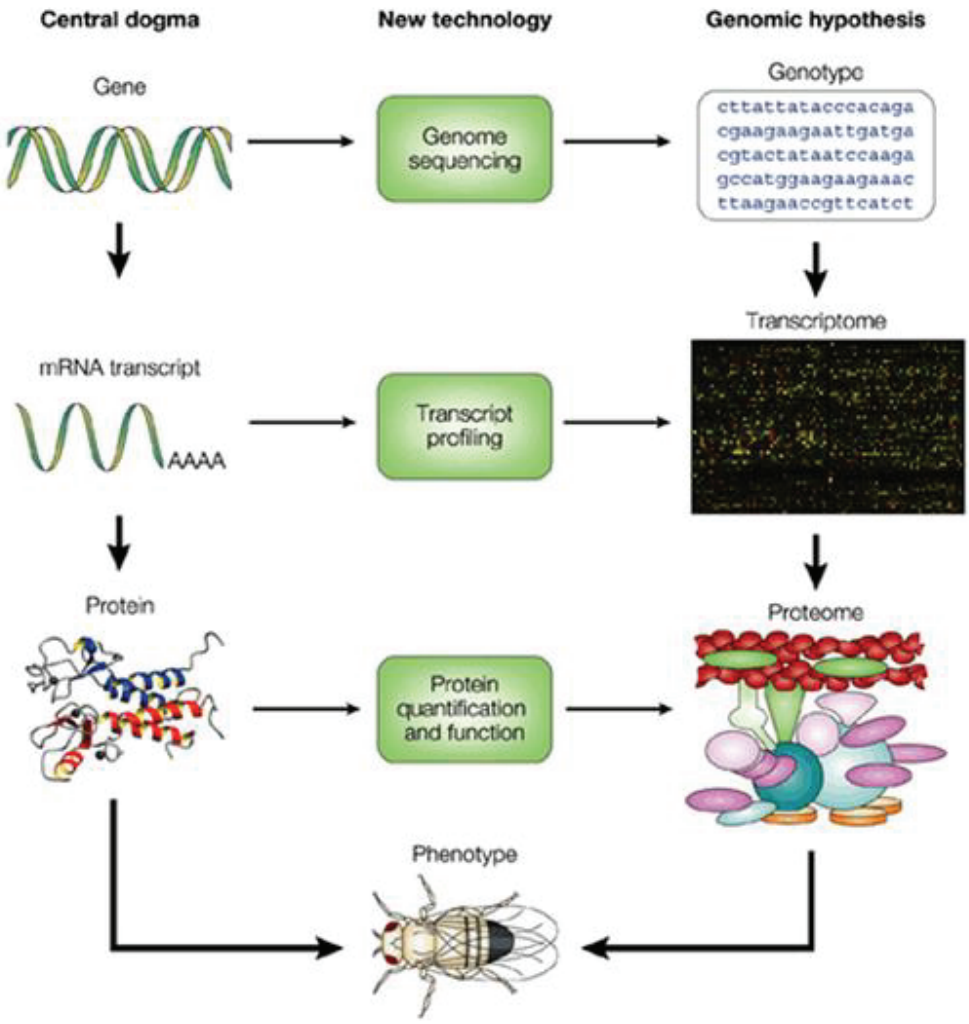
Figure 1.1 illustrates the central dogma in molecular biology (Doerge, 2002). In the organism, genetic information is stored as deoxyribonucleic acid (DNA). The genotype may be determined by sequencing the genome, the collection of all DNA within the organism. Parts of DNA are transcribed into ribonucleic acid (RNA), and a subset of RNA, the messenger RNA (mRNA) molecules, can be probed by techniques such as microarray or RNA-sequencing, for instance. These technologies provide information regarding the transcriptome, or the collection of RNA transcripts with which proteins will be produced. At the third stage in the central dogma, proteins translated from RNA transcripts can be detected and identified through, for example, mass spectrometry-based proteomic analysis. The proteome, or set of proteins within an organism, may be characterised in this manner.

1.3 Gene and Genome Annotation

The main objective of a single gene annotation is to describe its function (e.g. protein-coding) (Mudge et al., 2013). From the function of one sequence, the functions of other sequences that are homologous may be inferred, allowing more directed and efficient investigations of the latter sequences. In addition, sequence homology amongst different species (e.g. *Saccharomyces cereviae* and *Homo sapien*) enrich our biological understanding of what is genetically encoded by even distantly related organisms and when those species diverged in evolutionary history.

Historically, transcript models were developed on libraries created by Sanger-sequencing of cDNAs or expressed sequence tags (ESTs) (Conesa et al., 2016). More

Figure 1.1: A schematic diagram of the central dogma and associated recent technologies for analysis at each stage. Reproduced with permissions from Doerge (2002).



recently, RNA-sequencing has been more widely used for transcript annotation. For protein annotation, protein sequences from the Swiss-Prot database (Boutet et al., 2016) and *ab initio* ORF-finding methods were used to infer coding sequences (CDSs) (Stanke and Waack, 2003; Burge and Karlin, 1997).

A collection of transcript models is considered a 'genebuild' by GENCODE and Ensembl (Mudge et al., 2013). While a genebuild manually annotated by teams of curators is considered the gold standard, most of the new releases are produced computationally. *In silico* annotation has three main steps: 1) transcript alignment, 2) annotate by comparing CDSs to genomes of other closely-related species, and 3) predict annotations based on the likelihood of a sequence to code for a gene feature with *ab initio* programs such as AUGUSTUS Stanke and Waack (2003).

Although a genome may be considered well-annotated, missing gene features should still be detected and further investigated. Recently, new classes of RNA molecules have been discovered, such as microRNAs which have various regulatory functions (Lee et al., 1993; Wightman et al., 1993) and promotor-associated short RNAs which may elucidate pervasive RNA transcription observed in RNA-seq experiments (Kapranov et al., 2007). Therefore, not only are new individual genes discovered that elucidate cellular and molecular pathways and interactions, but entirely new classes of gene features are able to be characterised, expanding our understanding of genomics.

1.4 RNA-sequencing

High-throughput RNA-sequencing is one type of technology that describes and quantifies the transcriptome of an organism in a particular stage of development of physiological condition (Wang et al., 2009). Common goals of transcriptomics include the following:

- the characterization of all expressed members of a particular type of transcript (e.g. small RNAs, non-coding RNAs, and mRNAs)
- determine where transcripts start (5' end) and terminate (3' end), patterns of splicing, and where chemical modifications or molecular binding sites occur
- quantify any changes in levels of expression of transcripts under different developmental or physiological conditions

1.4.1 Previous Technologies

Below are descriptions of major previous technologies that preceded and led to the development of RNA-seq included hybridisation and sequence-based approaches.

Hybridisation Approaches

Microarrays involves a reverse transcription step from sample RNA to cDNA. Reverse transcriptase is used to convert mRNA into cDNA molecules tagged with fluorescent labels. The cDNA is then incubated with custom-made or high-density oligo microarrays, comprised of a collection of target DNA sequences that the cDNAs can hybridise against, bound to a glass slide. Hybridisation is then detected by fluorescence intensity (Wang et al., 2009). Some microarrays can be specialised

to detect isoforms that are spliced by using probes that span exon junctions (Clark et al., 2002). One advantage of hybridsation techniques is that they are relatively inexpensive. However, the methods rely on prior knowledge of the genome for designing the complementary sequence probes. In addition, cross-hybridisation causes a lower signal-to-noise ratio, making it more difficult to determine a true hybridisation event. Microarrays also have relatively small dynamic ranges of detection due to the cross-hybridisation and signal saturation. Often, complex normalisation methods are required for quantifying expression level changes, which can be difficult (Wang et al., 2009).

Sequence-Based Approaches

In contrast to hybridisation techniques, sequence-based approaches directly determine the sequences of cDNA. Expressed sequence tag (EST) sequencing by Sanger sequencing is one such approach (Wang et al., 2009). The disadvantages of using EST sequencing are that it is a relatively expensive technology, not quantitative, and low throughput.

There are also tag-based methods, including serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE), and massively parallel signature sequencing (MPSS). Although the high-throughput nature and digital output of transcript levels in tag-based methods are advantages over EST sequencing, these methods still rely on expensive Sanger sequencing technology. Moreover, a significant number of the sequence tags cannot be mapped uniquely to the genome. Only part of the transcript is analysed; therefore, isoforms are not easily distinguished.

1.4.2 Short-Read RNA-sequencing Technology

In general, short-read RNA-sequencing provides the relative abundance of the collection of transcribed RNA molecules present at a given moment within a sample.

Briefly, total RNA are extracted from the biological sample and enriched for mRNA using oligo-dT attached magnetic beads. RNA is then fragmented and primed for cDNA synthesis. The first strand cDNA is synthesised via reverse transcription and using random primers. The second strand cDNA is generated by separating the RNA template from the first strand via AMPure XP beads and synthesising a complementary strand to make ds cDNA. Fragmentation can create overhangs, so these ends are blunted using 3' to 5' exonuclease activity to remove the overhang. The 5' overhangs are filled in via polymerase activity. To prevent blunted 3' ends of fragments from ligating to each other during adaptor ligation, a single adenine nucleotide is added. In addition, a complementary thymine is added to create an overhang for adapter ligation. Indexing adapters are then ligated to ds cDNA ends in preparation for hybridisation. DNA fragments that containing adapters on both ends are selectively amplified via PCR. Quantification of the DNA library templates should be determined and quality control analysis performed. To prepare the DNA templates for cluster generation, indexed DNA libraries are normalised and pooled (if applicable).

1.4.3 Advantages of Short-Read RNA-seq

One advantage of RNA-sequencing over previous technologies is that prior knowledge of the reference genome and reference transcriptome is not necessary, which

makes it potentially useful for non-model organisms that do not have fully sequenced genomes. Short reads (about 30 bp) can provide information regarding which exons are spliced together, while longer or paired-end reads can be used to describe how several exons are connected. These properties of RNA-seq produced data render the method conducive for studying isoforms and other variations in transcript sequences like single-nucleotide polymorphisms (SNPs) (Cloonan et al., 2008; Morin, Ryan, 2008). There is no upper limit of quantification since the measurement is just the total number of sequences acquired, whereas quantifying transcripts with very low or very high abundances is considerably more difficult with microarrays. In RNA-seq, cloning is not required, and, in some instances, neither is amplification. A much smaller sample of RNA is needed, reducing the cost of RNA-seq experiments (Wang et al., 2009).

1.4.4 Challenges and Considerations in Short-Read RNA-seq

Challenges still persist when invoking RNA-seq and, thus, careful consideration needs to be given to the design and analysis of RNA-seq experiments to mitigate these.

Transcript End

Fragmentation of larger RNA molecules (around 200-500 bp) for sequencing can cause bias due to the reduction in RNA-seq signals for 5' and 3' ends of transcripts (Mortazavi et al., 2008). In addition, the fragmentation of cDNA molecules is strongly favorable toward the sequencing of 3' ends of transcripts (Nagalakshmi

et al., 2008).

Library Construction

The construction of the library is also a major consideration. Multiple copies of a single short read may appear in an amplified cDNA library, which could be a true representation of the abundance of the corresponding RNA molecule or a polymerase chain reaction (PCR) artefact (Sayols et al., 2016). This may be due to, for example, overloading of a flow cell that produces optical duplicates. One possible solution is to examine this behavior across multiple biological replicates.

Strand Specificity

The RNA-seq method may not provide information about which DNA strand the RNA molecule was derived from, yielding unstranded RNA-seq reads making it impossible to resolve reads from overlapping genes on opposite strands. Conversely, the method could give strand information but may require extensive preparation (Cloonan et al., 2008) or inefficient direct RNA-RNA ligation (Lister et al., 2008).

Read Alignment

Short reads can be mapped directly to the reference genome or assembled into contigs first before mapping. The Trinity platform assembles transcriptomes *de novo* in non-model organisms (Haas et al., 2013). One clear advantage is not needing a reference genome before analysing the RNA-seq reads. However, limitations include the creation of erroneous chimeras between isoforms or paralogs, especially if the RNA-seq reads are relatively short. Additionally, isoform misalignment may create

artificial polymorphisms.

Mapping reads to complex genomes with extensive alternative splicing can be more difficult as some reads will span splice junctions. One way to alleviate this issue is to create a separate library of junction sequences and map junction reads to this library (Wang et al., 2009). One advantage of using *S. cerevisiae* as a model organism in this study is that splicing is rare; therefore, mapping RNA-seq reads to splice junctions should not prove to be a major issue.

Multi-Mapping Reads

Multi-mapping reads, or reads that map to multiple genomic locations, can further complicate read alignment in RNA-seq analysis. One way to handle multi-mapping reads is to observe the number of reads aligned to neighbouring unique sequences, and then distribute the multi-mapping reads proportionately accordingly (Mortazavi et al., 2008; Cloonan et al., 2008).

Long repetitive genomic regions and multi-mapping reads with greater than about 100 copies each can complicate analysis even more. The acquisition of longer reads may help with these complications since there will be a higher probability of including a unique non-repetitive region to help locate where in the reference genome an RNA transcript was transcribed from (Wang et al., 2009). One method of producing longer reads is paired-end sequencing.

Sequencing Errors and Polymorphisms

Errors from the sequencing instrument and polymorphisms may be complex issues to solve with organisms that do not have reference genomes. However, many software read alignment programs, such as TopHat2 (Kim et al., 2013) and STAR read alignment program (Dobin et al., 2013), have options to manage single-base differences. Higher sequencing coverage and comparison across biological replicates helps to resolve these complications.

GC Content and Bias

GC-content bias must be considered when processing raw data from sequencers, since fragments that are GC-rich and GC-poor are underrepresented (Risso et al., 2011). For a given DNA fragment length, the higher the GC content, the more thermodynamically stable it is due to favorable effects on base-stacking (Yakovchuk et al., 2006). Since GC content varies amongst individual DNA fragments within a DNA library, clonal expansion may not occur evenly for all fragments (Risso et al., 2011). Thus, GC content variation confounds comparisons of raw counts amongst genes within a lane on a flow cell in addition to comparisons amongst replicate lanes. Normalisation methods for both within-lane (e.g. regression, global-scaling, or full-quantile normalisation methods) and between-lane biases may alleviate the effects of GC content variation.

1.4.5 Direct RNA-sequencing

As mentioned in the previous section on Short-Read RNA-sequencing, conversion of RNA to cDNA is required before sequencing. The conversion step requires a

reverse transcriptase, which are error-prone and may produce low levels of cDNA (Roberts et al., 1989). Since it is the cDNA that is sequenced, DNA contamination from other sources may also be sequenced and included erroneously in the library. Reverse transcriptases also have template-switching activity, which creates artificial antisense transcripts, fuses transcripts, and shuffles exons (Houseley and Tollervey, 2010). Therefore, to circumvent the challenges involved in converting RNA to cDNA, direct RNA-sequencing was developed to eliminate the conversion to cDNA altogether (Ozsolak et al., 2009). The first step uses the *Escherichia coli* poly(A) polymerase I to produce a poly(A) tail on 3' ends of non-polyadenylated RNA molecules to ensure that no nucleotides would be artifically added to the 3' end of the original RNA molecules in subsequent steps. Sequencing is initiated at the 3' end by adding polymerase and deoxythymidine triphosphate (dTTP) to base-pair with the adenine nucleotides in the poly(A) tail. Then, a mix of fluorescent Virtual Terminator nucleotides (C, T, A, and G) are added, and those that were not incorporated are washed away. Images are taken, and the fluorescent and inhibitor moieties of the nucleotide that was incorporated are cleaved to allow subsequent cycles. Although there is no conversion to cDNA, direct RNA-sequencing has disadvantages regarding single base errors, such as deletions, insertions, and substitutions. After applying direct RNA-sequencing to *Saccharomyces cerevisiae*, the longest perfect match aligned read was 50 bp long, so this technology is optimised for relatively short RNA sequences (Ozsolak et al., 2009).

PacBio Sequencing

Yet another RNA-sequencing technique was developed by Pacific Biosciences, where longer RNA-sequencing reads are produced (Rhoads and Au, 2015). Each end of a target dsDNA molecule is ligated to a hairpin adaptor, creating a SMRTbell (single-molecule real-time). The SMRTbell is loaded onto a SMRT cell chip, where it can diffuse into a zero-mode waveguide sequencing unit. A polymerase can bind to one of the hairpin adaptors to start replication using four fluorescent nucleotides. Nucleotide binding to the polymerase creates a light pulse that reveals its identity. A series of these light pulses creates a 'movie' that describes the DNA sequence, producing a continuous long read (CLR). Read lengths above 60 kb have been reported. However, because the replication process is limited by the lifespan of the polymerase, longer sequences produce fewer CLRs, yielding lower accuracy. Since repeat regions within a reference genome make alignment of short reads more challenging, a major advantage of the PacBio sequencing technology is the production of long reads that have a higher probability of spanning regions also containing non-repeat regions for easier mapping. Some limitations of PacBio sequencing include failure of a polymerase to anchor or loading of multiple DNA molecules onto a zero-mode waveguide unit. These issues result in only 35,000-70,000 of the 150,000 wells on a SMRT cell to produce successful reads, a major decrease in efficiency. Moreover, the error rate of a CLR is about 11-15%, but this can be reduced by increasing the number of sequencing passes. Greater than 99% accuracy was generated by 15 passes (Rhoads and Au, 2015).

Comparably, Oxford Nanopore technology also produces read lengths of these

magnitudes (Madoui et al., 2015).

1.4.6 Software Used in RNA-seq Analysis

STAR Aligner

In this study the decision was made to use the STAR (Spliced Transcripts Alignment to a Reference) aligner because of its relative speed and flexibility (Dobin et al., 2013). STAR is an RNA-seq alignment algorithm for high-throughput long and short RNA-seq data to a reference genome. In this study, STAR is implemented for mapping the nearly half a billion 50-bp RNA-seq reads to the *Saccharomyces cerevisiae* reference genome. This allows, for example, downstream analysis of quantifying read counts per gene. Through this quantification, RNA-seq alignment profiles of un-annotated regions may be compared to annotated regions.

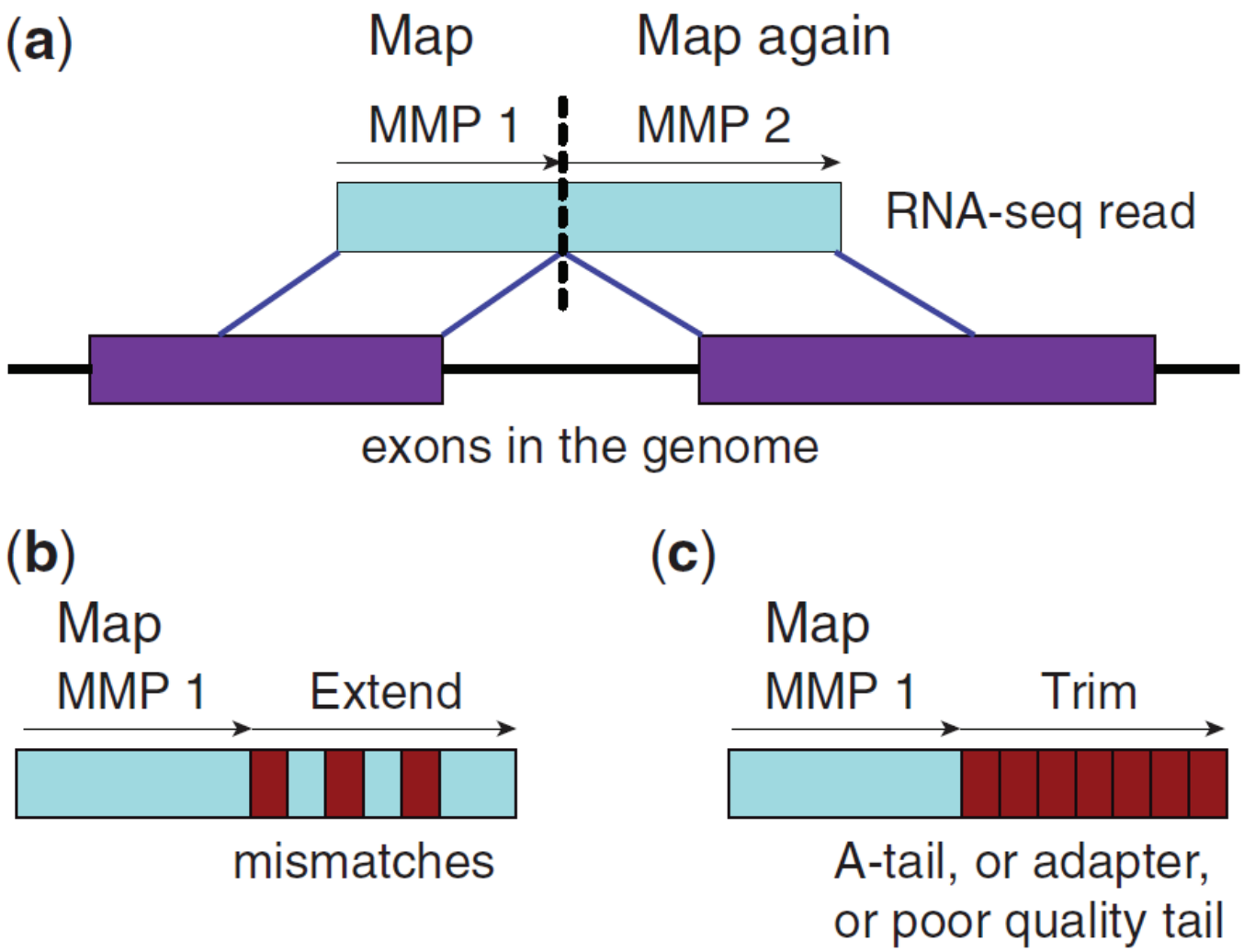
STAR operates through two phases: seed searching and clustering, stitching, and scoring (Dobin et al., 2013). In the seed search, the Maximal Mappable Prefix (*MMP*) is determined. The *MMP (R,i,G)* is the longest substring of read sequence R that matches at least one substring of genome sequence G exactly at read location i. Part (a) in Figure 1.2 illustrates aligning a read that spans a splice junction but does not contain any mismatches. Starting from the first base of the read, the STAR algorithm finds the *MMP*. Since the read spans a splice junction, the entire read cannot be contiguously mapped to the location on G. Therefore, the first portion (MMP1) is mapped to a donor splice site. For only the second portion of the read, the yet unmapped part, the *MMP* search is repeated. Because the search continues only for the unmapped portion, the algorithm is more efficient and quicker

than its competitors such as Mummer (Delcher et al., 1999, 2002; Kurtz et al., 2004) and MAUVE (Darling et al., 2004, 2010), which find all potential Maximal Exact Matches.

MMPs can also serve as anchors that can be extended when the end of a read cannot be reached due to mismatches (Figure 1.2 (b)). The extensions may allow for mismatches when parameters are set to permit them. Poor sequencing tails, poly-A tails, or library adapter sequences may be detected by the STAR algorithm if this extension step yields a low quality alignment (Figure 1.2 (c)). The *MMP* search is performed also in the reverse direction of the read and initiated at various points within the read through user-defined settings. This flexibility allows error detection at read ends and increases mapping sensitivity (Dobin et al., 2013).

In the second phase of clustering, stitching, and scoring, the seeds from the first phase are grouped and then connected. Proximity to a selected set of anchor seeds is used to cluster all other seeds (Dobin et al., 2013). Seeds that map within the genomic windows around anchors are stitched together in a linear fashion. In stitching each pair of seeds, any number of mismatches is allowed but only one insertion or deletion is permitted. A local alignment scoring scheme with user-defined penalties for mismatches and indels guides the stitching process to construct a quantitative analysis of read alignment qualities.

Figure 1.2: A schematic of diagram of how the STAR aligner searches for the Maximum Mappable Prefix while detecting (a) splice junctions, (b) mismatches, and (c) tails. Reproduced with permissions from (Dobin et al., 2013).



Heuristic Nature of Aligners and Sub-Optimal Mappings

Developers of short-read aligners, such as Burrows-Wheeler Alignment (BWA) (Li and Durbin, 2009) realised the limitations of their programs in terms of ability to accurately and quickly align longer reads that were produced by more advanced sequencing technology. BWA maps relatively short nucleotide sequences to long sequences, such as a reference genome by querying sequences that are shorter than 200 bp to perform gapped alignments. However, because the algorithm's stringency requires the entire read to be aligned, reads longer than 200 bp may be interrupted by structural differences, causing the program to fail. As a result, a newer implementation called Burrows-Wheeler Alignment Smith-Waterman (BWA-SW) was developed with heuristic capabilities that allow for faster alignment and for gapped alignment of reads up to 100 kbp (Li and Homer, 2010). A heuristic method does not guarantee an optimal alignment according to the scoring scheme because the method cannot determine whether a local optimum is the global optimum, which may not even exist. Since it is not feasible to check every possible solution due to the nature of the large amount of data, heuristic methods are invoked to estimate the global optimum. Many RNA-seq alignment programs followed a similar approach, including the more recent STAR aligner, which operates heuristically (Gingeras Group, 2016; Dobin et al., 2013). The program's algorithm searches for the maximum mappable length of a read by a split/search/extend method. Although the alignment program works for longer reads and take less computational time, STAR does not find all possible alignments since it does not perform local searches like the Smith-Waterman algorithm (Smith and Waterman, 1981). Because not all possible alignments are

found for each read, sub-optimal mappings may result, one disadvantage of using heuristic algorithms.

1.4.7 Sources of Variability and Error

Biological Variability

Biological samples under identical experimental treatments will exhibit biological variability. This manifests as different levels of RNA expression for the same genomic feature in RNA-seq. For instance, one population of cells may synthesise 10 copies of the protein Flocculin-1 (Chapter 3) per cell, while a second population may synthesise 30 copies of Flocculin-1, when only 5 copies are necessary for the same level of function. Having only 2 populations of cells, as in the aforementioned case, would create a mean of 20 copies per cell; however, if a higher number of populations of cells are used to calculate an average, for example, 10, 30, 5, 7, 9, and 4 copies per cell, then the mean would be about 11 copies per cell, which is a more representative estimate of the number of copies of Flocculin-1 that are produced by the majority of cells. Therefore, it is critical there is an appropriate number of biological replicates is determined for any experiment, taking into consideration the statistical power required for proper analysis (Conesa et al., 2016; Schurch et al., 2016).

Preparation of Sample

When extracting total RNA from a biological sample, mRNA can be isolated in two main ways: poly-A enrichment or ribosomal RNA depletion. One instance where ribosomal RNA depletion may be required is for formalin-fixed and paraffin-embedded samples, in which RNAs are degraded to smaller fragments that may not

still contain a poly-A tail for poly-A enrichment (Mullins et al., 2007). The Ribo-Zero-Seq protocol reduces 5' to 3' bias (poly-A capture methods yield more reads from the 3' ends of sequences) and has more uniform gene coverage; however, it has a higher detection rate of pre-mRNAs, resulting in fewer total reads that align to exon regions. Therefore, for the same transcriptome coverage, it was found that 2-4 times more reads had to be sequenced with the ribosomal RNA depleted library than its poly-A capture counterpart.

As an internal control, RNA-seq spike-in mixes from External RNA Controls Consortium (ERCC) can be included in RNA-seq samples to help quantify levels of expression. In one study comparing a poly(A) enrichment method to the RiboZero approach, it was found that ERCC-116 was underexpressed by 7.3-fold in the former (Qing et al., 2013). In a separate study, ERCC-116 was observed at concentrations that were 10 times higher or lower than expected (Seqc/Maqc-Iii Consortium, 2014). Consequently, it is imperative that the quality of internal controls such as spike-ins is carefully monitored and continually assessed throughout the experiment as low quality sequences may introduce bias in expression levels.

In the laboratory, other sources of DNA may contaminate RNA-seq samples, such as human skin, bacteria, viruses, other yeasts, and mycoplasma (Olarerin-George and Hogenesch, 2015). Consequently, some transcripts in the target organism may artificially appear to have higher levels of expression due to contamination from other organisms that share some homology. If preparing a cDNA library is required in the protocol, PCR may be another source of error since duplicate RNA-seq reads may be a result of clonal copies of a single transcript instead of actually having multiple RNA transcripts in the sample (Quail et al., 2012).

Normalisation

If normalisation methods are used in processing raw sequencing data, precautions should be taken to minimise the introduction of bias. If the lengths of genes are not considered for normalisation techniques that rely on total read counts, then normalised read counts may mask shorter genes that have lower total read counts due to its short length but may actually have higher read counts per base-pair (Bullard et al., 2010). Global normalisation methods may rely on total lane counts (commonly referred to as RPKM), per-lane counts (uses internal control genes that should be expressed consistently across biological conditions), or per-lane upper-quartile of gene counts (genes must have reads in one or more lanes) (Mortazavi et al., 2008).

1.4.8 Variability from Various Alignment Tools

Comparisons of Spliced-Alignment Programs

In a study done by Engstrom et al. in 2013, 26 different mapping methods, created from the programmes BAGET (MasonLab, 2016), GEM (Marco-Sola et al., 2012), GSNAP (Wu and Nacu, 2010), GSTRUCT, MapSplice (Wang et al., 2010), PALMapper (Jean et al., 2010), PASS (Campagna et al., 2009), ReadsMap, SMALT, STAR (Dobin et al., 2013), TopHat1 (Trapnell et al., 2009), and TopHat2 (Kim et al., 2013), were compared (Engstrm et al., 2013). The comparisons were based on the following parameters:

- Alignment yield
- Mismatches

- Basewise accuracy
- Indel frequency and accuracy
- Positioning of mismatches and gaps in reads
- Coverage of annotated genes
- Spliced alignment quality

Of all of the methods, GSNAP, GSTRUCT, MapSplice, and STAR were the highest performing alignment methods; however, these programs still have their disadvantages. The outputs of GSNAP, GSTRUCT, and STAR contain many false exon junctions, so junctions should be filtered on the number of supporting alignments. The study also noted that significant improvements were observed for alignment methods which used gene annotation, especially in terms of mapping spliced reads (Engstrm et al., 2013).

Comparison of Aligners with Reference Genomes for *Saccharomyces cerevisiae*

In Nookaew et al. 2012, Stampy (Lunter and Goodson, 2011), GSNAP (Wu and Nacu, 2010), and TopHat1 (Trapnell et al., 2009) were compared (Nookaew et al., 2012). Although Stampy produced the highest mapping accuracy for ORFs that contained high genetic variation, it took the most computational time. GSNAP required a shorter amount of time but resulted in the lowest mapping accuracy, which would be beneficial for analyzing large numbers of reads using genomes with minimal polymorphisms. In terms of accuracy and speed, TopHat was in between Stampy and GSNAP; however, the method did align reads on small exons well (Nookaew et al., 2012).

Differential Gene Expression

A comparison of baySeq (Hardcastle and Kelly, 2010), Cuffdiff (Trapnell et al., 2010), DESeq (Anders and Huber, 2010), edgeR (Robinson et al., 2010), and NOISeq (Tarazona et al., 2011) was also performed for differential gene expression analysis in S. cerevisiae under 2 conditions: respiro-fermentative (batch) metabolism and full respiratory (chemostat) metabolism (Nookaew et al., 2012). Each method was compared against results from the Stampy alignment program. There was overall agreement amongst all 5 programs since 33% of all differentially-expressed genes (DEGs) were commonly detected. EdgeR identified an additional 299 DEGs, more than any of the other individual methods (Nookaew et al., 2012).

Sequencing Quality Control Projects Comprehensive Assessment

Although there have been a number of comparisons done for RNA-seq alignment methods, including those mentioned previously, an extensive study compared performance and reproducibility amongst different RNA-seq platforms (Illumina HiSeq, Life Technologies SOLiD, and Roche 454) at multiple laboratories (Seqc/Maqc-Iii Consortium, 2014). With respect to differential expression and junction discovery, results from the 3 different RNA-seq platforms did coincide and agree overall. About 44,000 genes in human, 79% of those known, and about 310,000 exons, 47% of known ones, were detected across pairs of replicate sequencing sites. However, for a large number of genes, there were major deviations in the absolute expression levels amongst the platforms. For instance, 5,056 genes (9% of the total number of known genes) were observed with the HiSeq 2000 but not with the SOLiD technology. The deviations were noted to be systematic, and, therefore, not due to the lack

of reproducibility, but perhaps due to the accuracy of measuring absolute levels of gene expression (Seqc/Maqc-Iii Consortium, 2014). Although the exact cause(s) of these deviations were not able to be determined in that study, it was observed that the deviations were reduced in protocols that did not depend on poly-A selection. Additionally, effects of GC content and unfolding of sequence regions were not statistically significant contributing factors to deviations in absolute expression levels of genes.

1.4.9 Efficacy of Short-Read Mapping on Short Genes In Yeast

Splice Site Prevalence

Less than 5% of protein-coding genes in yeast contain introns (Parenteau et al., 2008); however, in organisms with large proportions of genes that are spliced, and in particular where there are many isoforms of each gene, such as Arabidopsis thaliana or human, having longer RNA-seq reads is a significant advantage as reads would have higher probabilities of overlapping, and thus providing support for, splice junctions. Since splicing events are relatively rare in S. cerevisiae, longer reads would provide a smaller benefit.

Exon Length

The mean length of protein-coding genes in the SGD *S. cerevisiae* annotation is calculated to be 1,764 bp (2). Mean lengths for promoter, 5 UTR, 3 UTR, and terminator regions are reported to be 455 bp, 83 bp, 136 bp, and 275 bp, respectively

(Tuller et al., 2009). Thus, a rough estimate of the mean exon length is about 815 bp. The set of RNA-seq reads are 50-bp and single-ended. To have an entire exon covered by at least one RNA-seq read and to have consecutive RNA-seq reads overlap on at least 1 base, fewer than 20 50-bp RNA-seq reads are required for a 815-bp exon. There were nearly half a billion reads available in this study and about 6,000 protein-coding genes in yeast. Because of the large number of reads available, and the relatively small number of and short lengths of protein-coding genes, short-read mapping does not pose a major issue in this study.

1.5 Proteomics

The use of mass spectrometry in proteomics allows for the identification and quantification of peptides and proteins within a sample (Aebersold and Mann, 2003). Mass spectrometry involves detecting and measuring the number of ionised analytes in the gas phase. Figure 1.3 illustrates a generic mass spectrometry experiment in which:

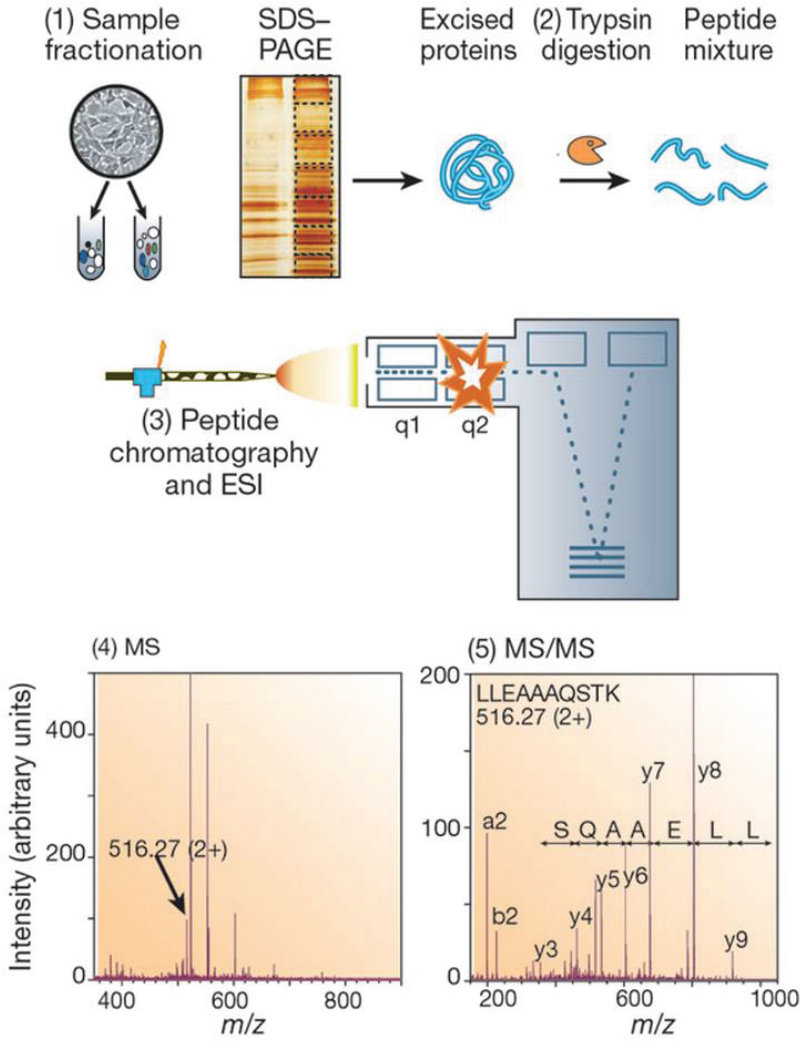
1. proteins are extracted and isolated from the cell, tissue, or organism and separated by SDS-PAGE

2. proteins are enzymatically digested into peptides by trypsin

3. peptides are separated by high-pressure liquid chromatography and nebulised into small charged droplets by electrospray ionisation

4. after peptides enter the mass spectrometer, mass spectra are taken at successive time points

Figure 1.3: A schematic diagram of a generic mass spectrometry method in proteomics. Reproduced with permission from Aebersold and Mann (2003).



5. a computer-generated prioritised list determines which peptides to fragment, and a series of tandem mass spec (MS/MS) experiments proceed

1.5.1 Mass Spectrometry Experiment Components

As shown in Figure 1.3, there are three main components in a proteomics mass spectrometry experiment: an ion source, a mass analyser, and a detector (Aebersold and Mann, 2003). The sections below describe common methods used in each

component of an experiment.

Ion Source

Electrospray ionisation (ESI) ionizes solutions to isolate analytes through nebulization(Aebersold and Mann, 2003). Matrix-assisted laser desorption/ionisation (MALDI), which uses laser pulses to sublimate and ionise analytes from a dry, crystalline medium. ESI can isolate constituents of complex mixtures, whereas MALDI is effective for relatively simple peptide samples (Aebersold and Mann, 2003).

Mass Analyser

The basic types of mass analysers are Fourier transform ion cyclotron (FT-MS), ion trap, time-of-flight (TOF), and quadrupole (Aebersold and Mann, 2003). Analysers can be stand-alone or arranged in tandem.

Ion traps capture ions for a certain amount of time and then allow ions to be analysed by mass spec. This method is relatively inexpensive, sensitive, and robust; however, ion traps have relatively low mass accuracy. There have been developments to improve ion traps, including linear or two-dimensional ion traps, which contain the ions in a much larger volume than the original three-dimensional ion traps. The increase in volume improves the mass accuracy and resolution in addition to sensitivity (Hager, 2002; Schwartz et al., 2002).

Fourier transform mass spectrometers also trap, or capture, ions but in a magnetic field under vacuum (Aebersold and Mann, 2003). This method has high resolution, mass accuracy, dynamic range, and sensitivity (Marshall et al., 1998; Valaskovic et al., 1996; Martin et al., 2000; Lipton et al., 2002). In general, ion traps and

quadrupole mass analysers are coupled ESI to yield collision-induced (CID) spectra of fragments of precursor ions. TOF analysers are coupled to MALDI to determine masses of whole peptides (Aebersold and Goodlett, 2001). When identifying proteins from mass spectra, CID spectra provide not only the peptide mass but also the peptide sequence.

1.5.2 Data Acquisition and Analysis

Data Acquisition

During protein digestion and separation, some sample can be lost through excessive separation steps. There may be incomplete digestion by trypsin, leading to a smaller number of peptides detected (Nesvizhskii, 2010). In tandem MS/MS, the mass accuracy and resolution of MS analyser (resolution: several-500 ppm) determines the accuracy of the peptide ions charge state (Nesvizhskii, 2010). To achieve maximum accuracy, the instrument should be fine-tuned, the room temperature should be tightly controlled, and both internal and external (computational) calibration should be performed.

1.5.3 Protein Identification

After obtaining mass spectra, peptides and proteins are then identified from the spectra. There are three basic ways in utilising the spectra: peptide sequence tag approach, cross-correlation, and probability based matching (Aebersold and Mann, 2003). After the spectra are examined, a protein hit list is formed based on identified peptides. In the peptide sequence tag approach, the peptide's mass is used in

conjunction with an unambiguous, short sequence of residues from the peak pattern in the spectra. The origin of the peptide is determined by the short sequence probe (Mann and Wilm, 1994). In the cross-correlation method, theoretical mass spectra are built from peptide sequences in the database (Eng et al., 1994). The best match is then determined by the cross correlation, or overlap, of the actual and predicted mass spectra. In probability based matching, fragments from peptide sequences in the database are calculated and compared against observed peaks in mass spectra (Perkins et al., 1999). A score is then calculated for the match, which indicates the statistical significance of the match.

This study's proteomics analysis is performed with Comet, an MS/MS sequence database search tool that utilises the cross-correlation method of protein identification (Eng et al., 2013). Initially, protein sequences from the search database are scanned to find consecutive amino acids combinations that match the mass of the peptide within a mass tolerance range of usually $\pm 0.05\%$ (Eng et al., 1994). This search can accommodate modifications such as phosphorylation by altering the amino acid masses used in calculating the peptide mass. After an amino acid sequence that fits within the mass tolerance range is found, it is scored. A higher (better) score is achieved for the following conditions:

- higher numbers of predicted fragments that match observed ions in the spectrum
- higher abundance
- continuity of an ion series
- if an immonium ion for His, Tyr, Trp, Met, and Phe is observed in the spectrum

with the accompanying amino acid (if the amino acid is not present, the score is lowered)

Multiple prospective amino acid sequences that may match to a single spectrum are then ranked highest to lowest according to their scores. The top 500 amino acid sequences are compared with the experimental spectrum using a cross-correlation analysis. With this method, an artificial 'spectrum,' which contains predicted mass-to-charge ratios of fragment ions of given amino acid sequences, is reconstructed. Collision-induced dissociation creates two types of ions after fragmentation: type-b ions that have retained its charge on the N-terminus and type-y ions that have retained its charge on the C-terminus after undergoing H rearrangement. The magnitude of the predicted mass-to-charge ratios are assigned as follows:

- 50.0 for mass-to-charge ratios that match type-b ions and type-y ions
- 25.0 for ratios that are within ± 1 of the b and y ion values
- 10.0 for type-a ions (neutral loss of ammonia, water, or carbon monoxide) with ± 1 mass-to-charge ratios

The reconstructed 'spectrum' is then compared to the experimental tandem spectrum, which needs to be processed first. The precursor ion's mass-to-charge ratio is removed from the experimental spectrum so that the major contributor in calculating the cross-correlation function (below) is similarity between fragment ion patterns, not the precursor ion.

Cross-Correlation Function:

Two analytes with the same atomic arrangement but different stable-isotope composition can be distinguished in a mass spectrometer due to the difference in mass

(Aebersold and Mann, 2003). Therefore, incorporation of stable isotopes in proteins and peptides to be measured by mass spectrometers enhances quantitative proteomics. Figure 1.4 illustrates three techniques in the stable-isotope labelling of proteins described in the sections below.

Metabolic Stable-Isotope Labelling

Cells are grown in isotopically enriched (e.g. nitrogen-15 salts or carbon-13-labelled amino acids) or depleted media (Conrads et al., 2002). Cells then incorporate these heavy isotopes into proteins when they are synthesized. Differences in peptide masses vary according to amino acid composition between the light and heavy peptides. The proteomics data used in this study were derived from yeast cells that were labelled via the stable-isotope labelling by amino acids in cell culture (SILAC) method, a sub-type of metabolic stable-isotope labelling (Ong et al., 2002).

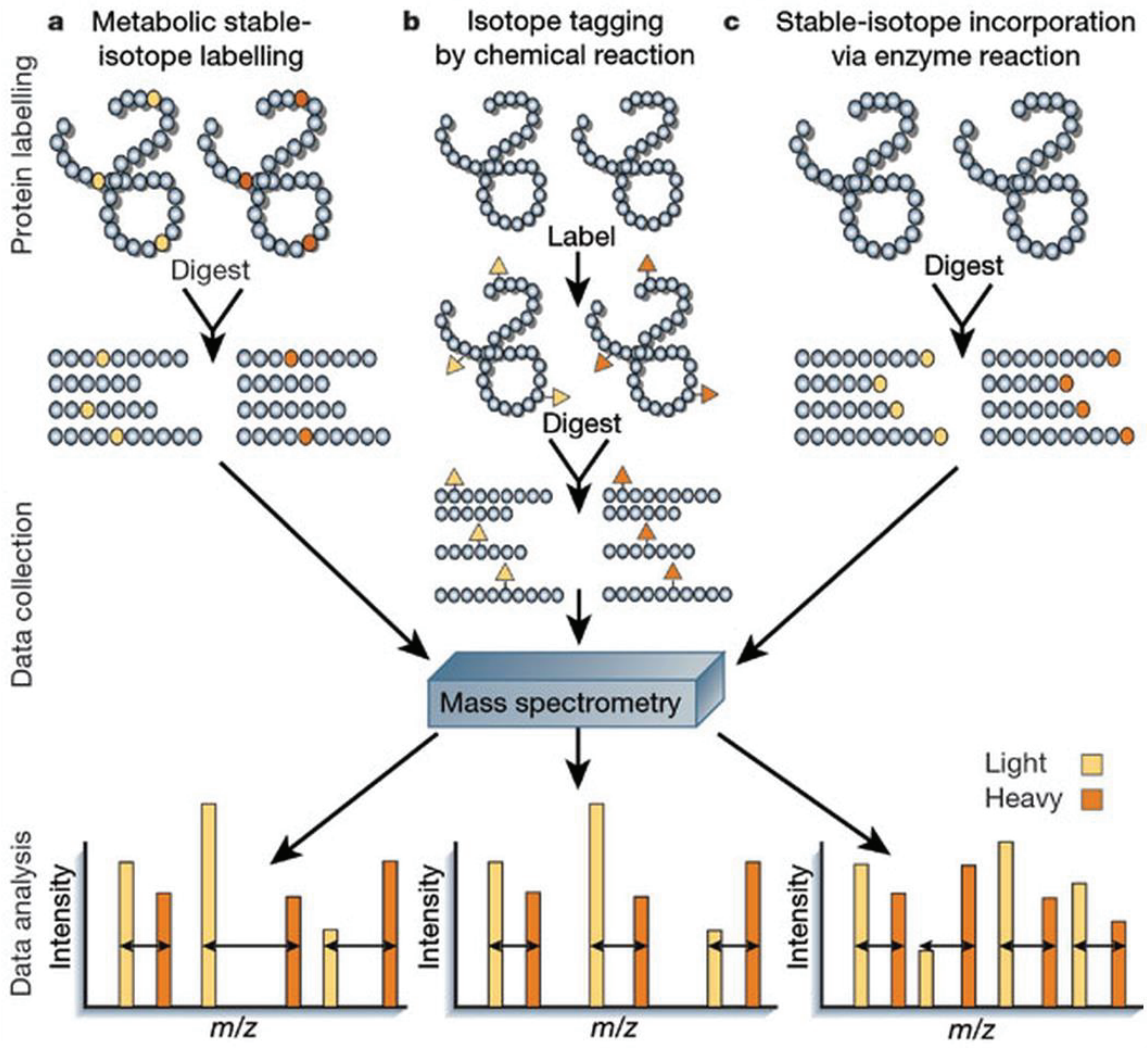
Isotope Tagging By Chemical Reaction

Proteins can be tagged by undergoing chemical reactions with reagents containing isotopic labels (Gygi et al., 1999; Zhou et al., 2002). If the reagents contain also affinity tags, after digestion peptides may be selectively isolated. The mass difference between peptide pairs is equal to a multiple of the mass difference of the isotopic label incorporated in the peptide by the reagent.

Stable-Isotope Incorporation Via Enzyme Reaction

During digestion, peptides can be labelled at the carboxy terminal with oxygen-18 from heavy water by an enzymatic reaction (Mirgorodskaya et al., 2000; Yao et al.,

Figure 1.4: A schematic diagram of three different methods of stable-isotope labelling in proteins. Reproduced from (Aebersold and Mann, 2003) with permissions.



2001). Quantitation is most difficult with this method out of the three since the mass difference either 4 or 2 Da.

False Discovery Rate

The target-decoy strategy can be used to calculate the false discovery rate (FDR), where decoys are reversed target sequences. The underlying assumption is that decoy PSMs and false matches to the target db have the same distribution (Elias and Gygi,

2007). There are two methods of incorporating the decoy database: searching the spectra separately against both the target and decoy databases or search against a combined target and decoy db. The latter is more commonly used. The FDR cut-off value for filtering PSMs can be estimated by two formulas:

- N_d/N_t = FDR (N_t = number of target PSMs with scores above cut-off, N_d = number of decoy PSMs among them estimated from N_{inc}, where N_{inc} = number of incorrect target PSMs). Assumption: the number of incorrect target PSMs can be estimated by the number of decoy PSMs because the number of target sequences is the same as the number of decoy sequences
- $2N_d/(N_t+N_d)$

One problem with separate searches is that since spectra that can be correctly matched to target seqs are allowed to match to decoys, Nd/Nt is a conservative estimate of FDR because Nd overestimates Ninc (incorrect) – correct by considering the ratio of incorrect PSMs in the whole dataset (Käll et al., 2008). The combined search is less sensitive to this problem with separate searches but is disadvantaged by the peptide competition effect. The peptide competition effect is when decoy PSMs have a higher score than the score of a true positive, which decreases the correct number of PSMs. There are two sources of incorrect PSMs:

- Truly random matches, which PSMs to reversed target seqs as decoys reflects well
- Incorrect PSMs to peptides that are homologous to the true peptides

This is a more severe problem at the level of protein-level identifications since it produces a bias in error rate estimation (Choi and Nesvizhskii, 2008; Feng et al.,

2007). There are currently no reliable methods to design decoys that mimic homology amongst sequences within a target database.

Data Variability and Error

One study quantified the amount of variability attributable to different aspects of the experimental process of using LC-MS/MS proteomics to perform a human brain tissue sample analysis (Piehowski et al., 2013). Variability was measured for differences in peptide abundance, by dividing the standard deviation of peptide profiles by the mean to calculate the coefficient of variation. Moreover, the ratio of peptide abundance to the median peptide abundance across all samples was used to calculated the global scaling normalisation coefficients. Protein extraction from tissue samples was by far the largest source of variation, accounting for 72% variation. Instrumental variance (fluctuation in instrumental response from one run to the next run in the short-term) contributed the second largest amount of variability overall at 16%, followed by instrumental stability (drift of the quantitation from the LC-MS/MS platform across the two weeks of continuous analysis) at 8.4%. The smallest amount of variability was derived from protein digestion at 3.1%. There are two primary causes of incorrect peptide identifications: random high-scoring matches of MS/MS spectra to unrelated sequences and matches to peptides homologous to true peptides (Nesvizhskii, 2014).

1.5.4 Proteogenomics: The Role of Proteomics in the Analysis of Genomes

Computationally, new potential protein-coding genes can be detected via sequence similarity or ab initio methods, the latter of which rely on properties such as codon usage, splice site consensus sequences, and GC-content (Ansong et al., 2008). Sequence similarity may be found by searching for homology in databases of known protein-coding genes with algorithms such as BLAST (Altschul et al., 1990) or searching for orthologs in COG (Tatusov et al., 1997) or PFAM (Bateman et al., 2004), for instance. However, it was estimated that for some eukaryotic genomes, only 50% of correct gene structures are predicted from de novo programs (Guigó et al., 2003). Even after receiving a highly evidence-supported prediction from a predictor, the potential candidate gene needs to be experimentally verified. One method is by invoking systematic RT-PCR to validate that the gene is indeed being transcribed (Wu et al., 2004); however, RT-PCR does not provide evidence that the transcript is translated. Because of this, the detection of peptides and proteins through proteomics is a powerful tool in helping to improve the curation of genomes. In addition, using isotopic labelling would build another layer of confidence if both unlabelled and labelled versions of the same peptide sequence were detected. Another advantage of using shotgun proteomics instead of more targeted sequencing approaches is that the database of sequences to which the spectra are searched, is entirely customisable. MS/MS spectra may be searched against any theoretical peptide sequence, facilitating efficiency and flexibility in investigating numerous potential protein-coding genes.

Expanded/Inflated Databases in Proteomics and Issues in Estimating FDRs

The larger the search database, the lower the sensitivity of peptide identification due to false positives. There is an increased chance in receiving a high scoring random match with increased db size (Nesvizhskii, 2014). This observation is an important consideration when choosing to search proteomics spectra against a database comprised of 6-frame translations of all ORFs within an organisms entire genome. The sequence nature and size of such databases underestimate the confidence scores assigned to peptide to spectrum matches (Blakeley et al., 2012). This effect is due the inflation of the decoy database: for every putative ORF, there are 5 other sequences that are incorrect with low probability of being protein-coding. Typically, decoy databases are created by reversing sequences in a target sequence database (e.g. known protein-coding genes). Therefore, in that instance, each correct sequence has only one incorrect sequence. Furthermore, false discovery rates and posterior error probabilities (which estimates the probability of a PSM being incorrect) calculated from proteomics searches on 6-frame translation databases result in fewer PSMs being accepted (Blakeley et al., 2012). These issues can be alleviated by selecting sequences from the database to yield higher-quality decoys, especially when the ratio of sequences in target and decoy databases are as close to 1:1 as possible. Based on these findings, although 6-frame translation databases were used for the proteomics searches presented here, filtering putative ORFs based on length dramatically reduced the number of sequences included (Chapter 4). In the most stringent filtering, the ratio of known protein-coding genes to potential protein-coding ORFs was very

close to 1:1.

1.5.5 Software Used In Proteomics Analysis

For the analysis presented here, the open-source and freely available Trans-Proteomic Pipeline (Deutsch et al., 2010) would be used for the proteomics analysis since it streamlines four major software programs : Comet (Eng et al., 2013), PeptideProphet (Keller et al., 2002), iProphet (Shteynberg et al., 2011), and ProteinProphet (Nesvizhskii et al., 2003), which are described in this section.

Comet

Comet is an open-source program for searching MS/MS sequence databases (Eng et al., 2013) which scores peptide sequences against observed spectra via an algorithm called fast cross-correlation. Comet has been developed from SEQUEST (Eng et al., 1994), its predecessor. With Comet, the creation and storage of theoretical spectra is not necessary due to an additional spectral pre-processing step, improving efficiency. The additional spectral pre-processing step, which results in both positive and negative intensity values, was developed after a mathematical rearrangement of the original cross-correlation function from SEQUEST (Eng et al., 2008). With the new fast cross-correlation algorithm, for each calculated fragment ion mass, the summation of peak intensities from pre-processed spectra is the cross-correlation score. The cross-correlation score is computed for each peptide in the database. A histogram of all cross-correlation scores is then log-transformed. A linear least-squares fit is applied to the underlying distribution of the transformed histogram. The cross-correlation score of the top-scoring peptide is projected down to intersect

the linear fit. The inverse log of the y-axis value of this projection is then considered the expectation value or E-value, which is an estimate of the number of peptides that are expected to score this well or better by chance.

PeptideProphet

After MS/MS spectra have been processed by Comet, the PeptideProphet (Keller et al., 2002) program implements an expectation maximisation algorithm to distinguish correct and incorrect peptide spectral matches (PSMs). The algorithm assigns each PSM an initial probability of correctness, based on machine learning from a training dataset, and a set of global false discovery rates are calculated as a function of the probability cutoff. Characteristics such as the number of missed cleavages, number of enzymatic termini, retention time, and mass deviation are considered to determine the qualities of PSMs. If data are produced from a high accuracy instrument, a high mass accuracy model may be applied, which models the deviation in the observed mass to the nearest isotopic peak (Keller et al., 2002). A target-decoy strategy can be employed in PeptideProphet in which models can be refined by user-defined target sequences (true peptides) and decoy sequences (false peptides). As final output, PeptideProphet produces a probability for each PSM in a pepXML file, modelling results, and receiver operating characteristic curves.

iProphet

When used in tandem with PeptideProphet, iProphet can integrate other characteristics of the entire set of PSMs, including multiple discoveries of the identical peptide

ion across different spectra, different charge states of the same peptide, or modifications of peptides to better reflect the nature of shotgun proteomic data (Shteynberg et al., 2011). Model refinement occurs with additional pieces of evidence.

One of the pieces of information is the number of replicate spectra (NRS), where the NRS score is positive for a precursor ion commonly identified at probabilities greater than 0.5, 0 if identified from only one spectrum, and negative if commonly identified at probabilities less than 0.5. This scoring system preserves the high probability of precursor ions that are identified only once.

The number of sibling searches (NSS) is a formula that sums all probabilities, calculated by PeptideProphet, from multiple search engines that agree on the same peptide sequence for one spectrum. The sum is then divided by the number of other searches performed on the spectrum. Hence, the NSS formula rewards identifications that have more consensus amongst search engines while penalising those that do not.

The number of sibling experiments (NSE) is a statistic that models the identification of one precursor ion across multiple experiments. If the precursor ion is commonly identified with probabilities above 0.5, the NSE yields a positive value, and vice versa. Precursor ions identified in only one experiment are assigned an NSE of 0.

If multiple precursors with different charges help identify a peptide, the peptide is rewarded in its number of sibling ions (NSI) statistic.

If different mass modifications contribute to the identification of a peptide, the peptide is rewarded in its number of sibling modifications (NSM) statistic.

More accurate posterior probabilities and global false discovery rates are calculated by iProphet after incorporating these aforementioned attributes of the proteomics dataset (Deutsch et al., 2010).

ProteinProphet

After statistical refinement by iProphet on peptide search results against MS/MS spectra, ProteinProphet is invoked to calculate the probabilities that proteins were present in a sample (Nesvizhskii et al., 2003). The primary input for ProteinProphet consists of a list of peptides assigned to MS/MS spectra with respective probabilities indicating the accuracy of the assignments. By calculating the number of unique peptides per protein, ProteinProphet increases the probability values of peptide assignments if there are multiple sibling peptides found and decreases probabilities of peptides for which no siblings were found. ProteinProphet then creates a list of proteins by collapsing redundant database entries into a single identification. Proteins that are indistinguishable based on peptide assignments are grouped. For each protein in the sequence database, a probability value is calculated to indicate the chance of its presence in the sample. The final output from ProteinProphet is a list of proteins with their respective peptide spectra match(es), protein probabilities, and global false discovery rates for various thresholds (Deutsch et al., 2010).

1.6 Sequence Analysis

Any new DNA, RNA, or peptide sequences found through this type of study can be classified and characterised according to similarities with other known sequences

from various databases. There are several current computational methods that help to search DNA and RNA sequences of interest against known information throughout databases. BLAST (Altschul et al., 1990) and its variants search the query sequence of interest against curated databases of nucleic and protein sequences based on sequence identity in alignments. Phylogenetic trees based on maximum likelihood describe relationships amongst sequences based on probabilities of base substitutions (Arthur Lesk, 2008). Conserved elements, or regions, within multiple sequence alignments can be found by using phylogenetic Hidden Markov Models (Siepel et al., 2005). InterPro (Apweiler et al., 2001) is an integration of several databases containing various types of information on proteins, which aims to determine relationships amongst proteins based on distinguishing signatures, or specific patterns of characterstics. Many of these software programs were implemented in conjunction throughout the study and thus described in this section.

1.6.1 Software Used in Sequence Analysis

BLAST and Its Variants

The Basic Local Alignment Search Tool (BLAST) queries a sequence, of amino acids (aa) for instance, against a database (Altschul et al., 1990). BLAST cuts a probe sequence into all possible k-mer segments, or words. For example, an amino acid sequence of interest may be cut into 4-aa long segments. For each of the 4-aa words, BLAST finds all occurrences of the word in the database, creating an index, which increases the efficiency of searching. At each occurrence, BLAST attempts to extend the matched region in the database entry in both directions, according to

Table 1.1: List of BLAST program (Altschul et al., 1990) names, functions, databases queried. Greater details on the databases are provided in Tables 1.2 and 1.3.

Program	**Function**	**Database Queried**
blastn	nucleotide query to search nucleotide databases	nucleotide collection (nr/nt)
blastp	protein query to search protein database	non-redundant protein sequences (nr)
blastx	translated nucleotide query to search protein databases	non-redundant protein sequences (nr)
tblastn	protein query to search translated nucleotide databases with BLAST	nucleotide collection (nr/nt)
tblastx	translated nucleotide query to search translated nucleotide databases	nucleotide collection (nr/nt)

the probe sequence, without allowing for mismatches or gaps. BLAST then merges these extended matches by aligning them, allowing mismatches and gaps. Scoring is based on gaps and the BLOSUM62 matrix, which gives penalties based on specific amino acid substitutions for mismatches.

In 1997, BLAST was refined to allow for gaps in extensions of the short word matches, called gapped BLAST (Altschul et al., 1997). In addition, the short word matches could be near-matches, instead of only exact matches as previously required. BLASTZ (Schwartz et al., 2003) is a modified version of gapped BLAST, in which matched regions must be in the same orientation and order. Additionally, different gap penalising and scoring methods are used which help prevent a gapped alignment from being triggered by biased nucleotide content.

Table 1.1 list the main BLAST programs and which databases are queried.

Table 1.2: The nr collection contains non-redundant protein sequences from the following databases.

Database	**Description**
GenPept (not an official release by NCBI)	Database of GenBank (Benson et al., 2013) gene products (translations of all coding sequence features). See Table 1.3 for GenBank description.
SwissProt (Boeckmann et al., 2005)	Extensively curated, manually annotated, and reviewed protein classifications and functions.
PIR (Wu et al., 2003)	Classified and functionally annotated protein sequences from the Atlas of Protein Sequence and Structure (Margaret O. Dayhoff, 1969).
PDF	Source and contents unknown.
NCBI RefSeq (Tatusova et al., 2014)	Non-redundant, well-annotated, comprehensive, and integrated database of DNA, transcript, and protein sequences.

Table 1.3: The nt collection consists of partially non-redundant nucleotide sequences from all traditional divisions of the following databases.

Database	**Description**
GenBank (Benson et al., 2013)	Collection of all publically available DNA sequences.
EMBL (Stoesser et al., 2002)	Nucleotide sequences from all available public sources are incorporated, organised, and distributed through this database.
DDBJ (Kosuge et al., 2014)	The DNA Data Bank of Japan is a collection of all freely available nucleotide sequence data. The nt database excludes the following divisions: GSS (genome sequence reads short single pass), STS (tag sites of sequences for genome sequencing), TPA (third party annotations), EST (expressed sequence tags, which are cDNA sequence reads short single pass), HTG (high throughput genomic sequences mainly from genome sequencing projects), and WGS (fragment sequences during whole genome shotgun assembling processes).

MULTIZ

Typical whole genome alignment strategies compare various genomes against a fixed reference genome (Blanchette et al. 2004). One disadvantage is that regions conserved in some species of interest but not present in the reference are not detected. Therefore, the program MULTIZ was built to align multiple sequences first and then project local alignments, or blocks, against one of the genomes of interest (Blanchette et al., 2004). Firstly, MULTIZ implements BLASTZ (Schwartz et al., 2003) to align sequences of interest in a pair-wise manner. Next with the Threaded Blockset Aligner (TBA) (Practical Guide to using TBA), a multiple sequence alignment (MSA) is generated from a given phylogeny tree depicting the relationships amongst the sequences and the pair-wise alignments. This step produces blocks, or continuous regions where one or more sequences align. It is possible for only one sequence to be present in a block if no other sequences contain regions that match. TBA can then project the whole MSA onto one of the sequences of interest, or reference. The projection is achieved by ordering all blocks that contain, or are threaded by, the reference sequence itself (Blanchette et al., 2004).

Phylogenetic Trees and Maximum Likelihood

A phylogenetic tree can be used to show relationships amongst sequences (Arthur Lesk, 2008). The lengths of branches connecting two sequences in the tree indicates how distantly related they are. Maximum likelihood is a method of creating a phylogenetic tree based on probabilities of base substitutions (the replacement of adenine, cytosine, guanine, or thymine in DNA with another base). For a set of sequences, all possible trees are considered. For each tree, substitution rates are

adjusted to produce the highest likelihood of yielding the actual sequences. The tree that gives the maximum likelihood is the optimal one.

Hidden Markov Models

A multiple sequence alignment (MSA) of a group of homologous sequences has patterns that define how they are related (Arthur Lesk, 2008). A Hidden Markov Model (HMM) is a computational structure that produces, or emits, a sequence based on the MSA's pattern statistics (Krogh et al., 1994). Figure 1.5 shows the structure of an example HMM that consists of states (circles) and choices (arrows). In this particular HMM, there are three types of actions available:

- D = deletion (starts or continues a gap)
- I = insertion (an amino acid is placed between two adjacent amino acids in the MSA)
- M = match (an amino acid is present in both the MSA and sequence emitted - not necessarily identical)

After entering the structure from the START, the choice of the next state is determined by the distribution of probabilities for the choice of the successor state (Arthur Lesk, 2008). Each state that emits an amino acid is also governed by another probability distribution for the 20 amino acids. These probability distributions are dependent on the particular group of related sequences the HMM describes. The subsequent state depends only on the current state, not previous states.

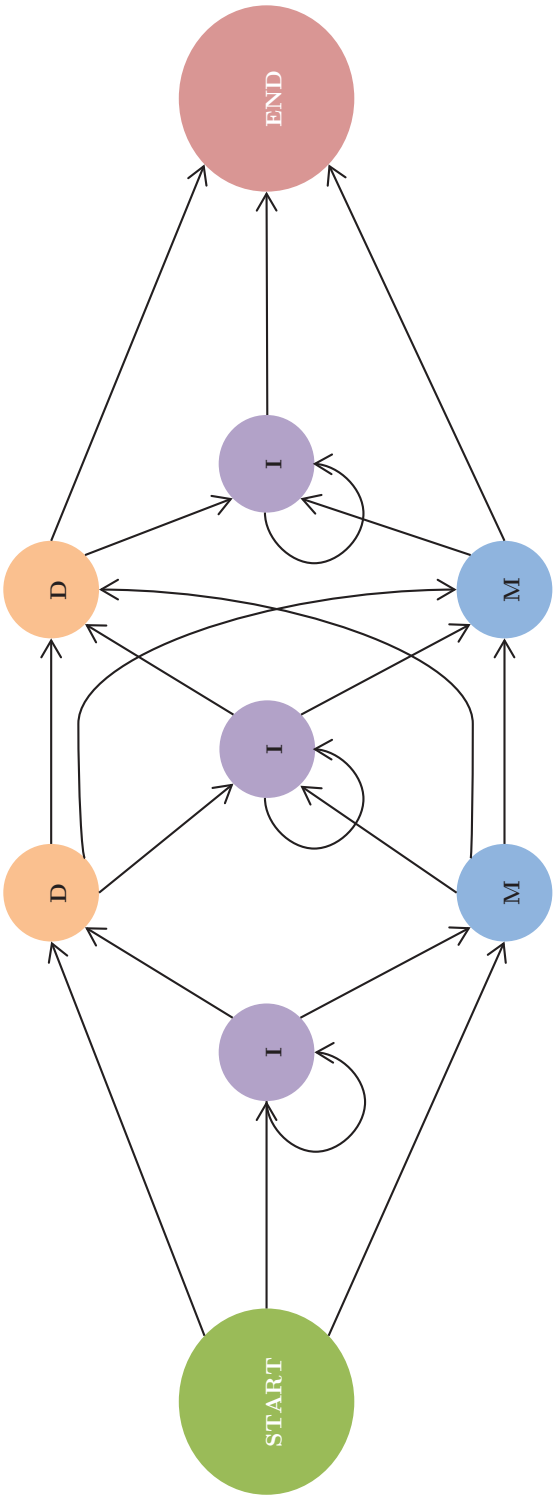


Figure 1.5: An example of a Hidden Markov Model that describes a multiple sequence alignment and emits an amino acid sequence. There are three states or nodes: D (deletion), I (insertion), and M (match). Probability distributions determine which residues are emitted at I and M states in addition to the successor state. Adapted from Krogh et al. (1994) with permission.

An example of a HMM application trained on a family of sequences is to determine whether a new sequence belongs to the family (Arthur Lesk, 2008). If the HMM can emit the new sequence with high probability, then the sequence is deemed as belonging.

phastCons

PhastCons is a program that predicts conserved elements in a multiple sequence alignment by invoking a phylogenetic Hidden Markov Model (phylo-HMM) (Siepel et al., 2005). The HMM consists of two states (nodes): conserved region and non-conserved region. Instead of a probability distribution, each state is governed by a phylogenetic tree built by maximum likelihood. Each branch on the phylogenetic tree represents a single or multiple substitutions (mutation events). The phylogenetic trees for both nodes are the same except that the branch lengths for the conserved node tree are scaled down to values between 0 and 1, yielding smaller distances between mutation events and therefore higher conservation. For each base in a MSA, the phastCons can compute a probability that a base was emitted by the conserved state in the phylo-HMM and thus conserved.

InterPro

Protein identifying information, or signatures, from several databases are systematically merged in the The InterPro database (Apweiler et al., 2001). The method of integrating signatures relies on characterising two hierarchical family relationships. The sub-string relationship exists amongst motifs that exist within a sequence region described by a wider pattern (e.g. a PROSITE pattern within a PRINTS

fingerprint). These signatures have identical InterPro entry accession numbers. On the other hand, the sub-type relationship describes motifs specific to a subset of sequences that are contained within another more general pattern, as in a parent-child relationship. Contrastingly, these signatures are given different accession numbers. Databases from which protein signatures were merged are listed in Table 1.4.

Table 1.4: List of InterPro member databases and their descriptions.

Database	**Description**
Pfam (Finn et al., 2014)	Curated database of protein families, each defined by two alignments and a profile Hidden Markov Model . A large sequence database, mainly constructed from the UniProt Knowledgebase (UniProtKB) (Consortium, 2012), is searched via the profile HMM for all member proteins in a family.
PRINTS (Attwood et al., 2012)	Consists of protein fingerprints, groups of conserved motifs that characterise a family of proteins. PRINTS iteratively scans the composite UniProtKB/Swiss-Prot-TrEMBL database (Consortium, 2012).
PROSITE (Sigrist et al., 2013)	Collection of patterns and profiles that identify protein families, domains, and functional sites. ProRule is a set of manually created rules to refine the differentiation of PROSITE motifs based on functionally and/or structurally critical residues. PROSITE, in conjunction with ProRule, annotates features and domains in entries of the UniProtKB/Swiss-Prot database.
ProDom (Bru et al., 2005)	Database of protein domain families, providing a multiple sequence alignment of homologous domains and a consensus sequence per entry. The database was automatically built by clustering homologous segments of non-fragmentary sequences from the UniProtKB/Swiss-Prot-TrEMBL database.
CATH-Gene3D (Sillitoe et al., 2015; Lees et al., 2012)	CATH is a method of classifying protein structures in the Protein Data Bank (PDB) (Bernstein et al., 1977), a database of atomic coordinates that describe protein structures from methods such as X-ray crystallography and NMR spectroscopy. Gene3D assigns CATH domain families to Ensembl (Flicek et al., 2014) and UniProt protein sequences that do not have PDB structures.
HAMAP (Pedruzzi et al., 2013)	Pipeline that automatically annotates proteins in UniProtKB using profiles of and manually created annotation rules for protein families with well-defined function and high sequence conservation.
PANTHER (Thomas et al., 2003)	Protein classification system according to families/subfamilies, biological process, and pathways.

PIRSF (Nikolskaya et al., 2007)	Method of classifying UniProtKB sequences in non-overlapping clusters based on characteristics of proteins as a whole, such as biochemical and biological functions, to depict evolutionary relationships.
SMART (Letunic et al., 2015)	Identifies and annotates protein domains according to tertiary structure, phylogeny, functional class, and functionally critical amino acids.
SUPERFAMILY (Wilson et al., 2009)	Database containing structural and functional annotations for proteins and genomes. SCOP (Andreeva et al., 2007) structural protein domains are described by a set of Hidden Markov Models. Protein sequences from over 2,478 completely sequenced genomes are searched against the HMMs to create SUPERFAMILY annotations.
TIGRFAMs (Haft et al., 2013)	Collection of curated multiple sequence alignments, protein sequence classification Hidden Markov Models, and other information for automatic annotation of prokaryotic proteins.

Chapter 2

RNA-seq and Un-Annotated Regions

Introduction

The RNA-seq dataset, consists of a high volume of data (∼400 million 50-bp reads) for a relatively small genome (∼12 million bp) (Cherry et al., 2012).

In this dataset we observed that some genomic regions that did not have an existing annotation were nevertheless highly expressed. This observation became the premise of this study, an attempt to investigate why transcription was active in un-annotated regions.

The development of the Un-Annotated Region Pipeline facilitated the analysis of these UARs with respect to the three RNA-seq alignment methods. Briefly, the UARs were sorted by total read count and conservation. The top UARs were then categorised based on the profiles of their RNA-seq read depths to find those with a distinctive, continuous region of high RNA-seq read depth. These regions

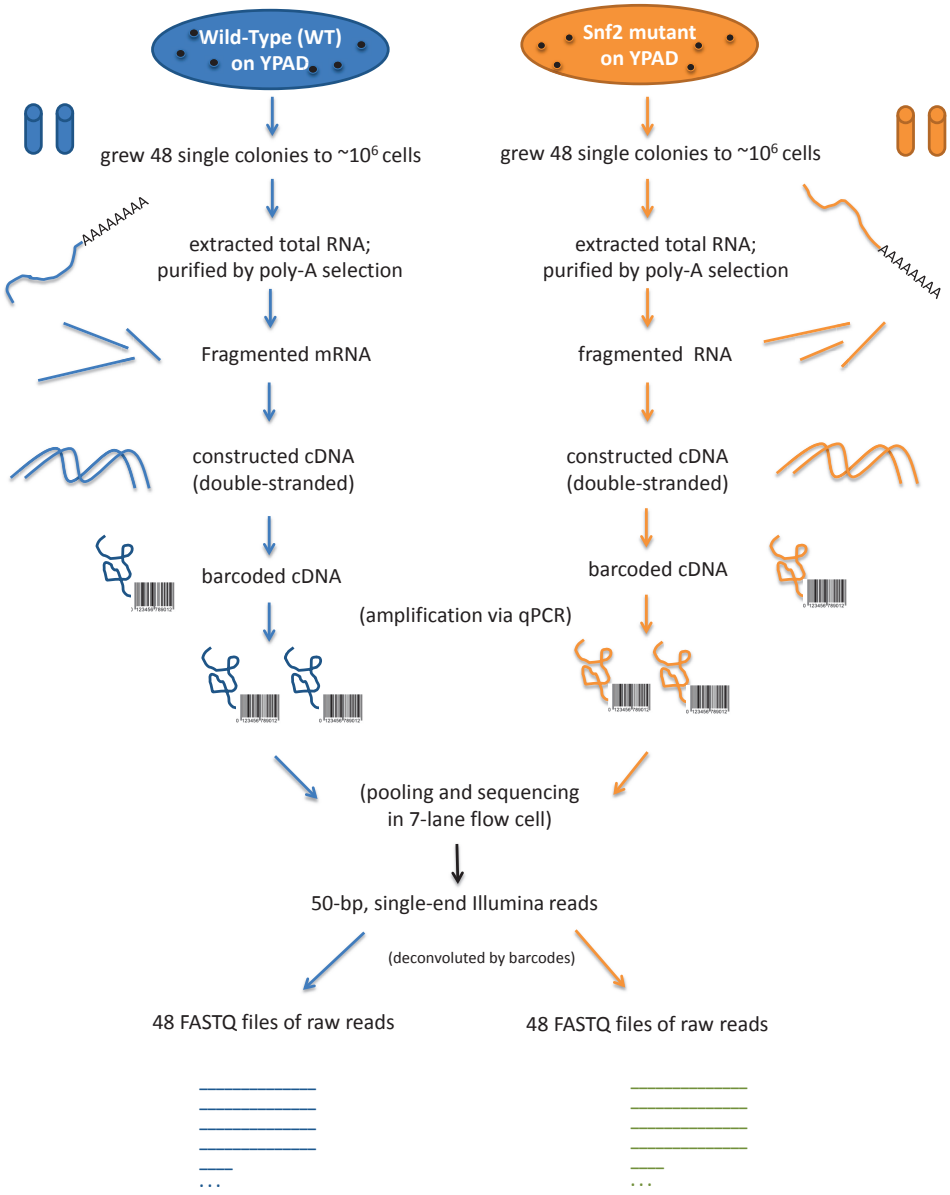
may indicate active transcription, and, therefore, the possibility of a new genomic feature.

2.1 The Experiment and Dataset

Figure 2.1 is a schematic of the methods used to culture the strain BY4741 S. cerevisiae, extract the RNA, and produce the RNA-seq data. Firstly, wild-type and $\Delta snf2$ mutant strains were grown on yeast extract-peptone-dextrose medium + adenine (YPAD) plates. This project was concerned only with the wild-type samples. For each plate, 48 single colonies were isolated and cultured in individual flasks and grown at 30 degrees C to a density of one million cells at OD600 of 0.7-0.8. Total RNA was extracted from each sample (culture) with the Qiagen RNeasy mini kit. Yeast cells were lysed with Zymolase, and DNA contamination was limited by DNase treatment. The range of total RNA extracted per sample was 30.3-126.9 micrograms. To ensure that distributions of RNA content were equal amongst the samples and biological conditions, the Kolmogorov-Smirnov test was performed and resulted in p=0.16, (Schurch et al., 2016). ERCC spike-in transcripts were then added to each sample as an internal control (Jiang et al., 2011; Lovn et al., 2012). The libraries were prepared via the standard Illumina multiplexed TruSeq method. The RNA samples were then purified by polyadenylation enrichment using poly-dT beads. After fragmentation, the first and second strand cDNA synthesis of RNA transcripts occurred, where cDNAs were then barcoded under the balanced block design to minimize technical artifacts and biases (Kaisers et al., 2014). Un-barcoded adapters were ligated from the cDNA sequences, and barcode-specific PCR

primers were used to enrich samples. The quality of libraries were assessed and passed. Samples were then diluted to 10 nM and underwent fluorescence-based quantification. The entire pooled library was separated into 7 individual pools, which were sequenced on an 8-lane flow cell with an Illumina HiSeq 2000, resulting in each lane containing sequences from each of the 96 total samples. After the flow-cell ran for 51 cycles single-end, this method produced a total of about half a billion 50-bp single-ended RNA-seq reads for each biological condition (wild-type and Δ*snf2* mutant). Barcodes for the reads were deconvoluted via the Illumina Cassava pipeline v1.8 to produce a single FASTQ file of the raw reads were produced for each of the 48 samples for each condition. The program fastQC was used to assess the quality of the reads, which were then mapped to the Ensembl release 68 genome annotations (Flicek et al., 2011) with bowtie2 (Langmead and Salzberg, 2012) and TopHat2 (Trapnell et al., 2009). Mapped reads were combined with htseq-count (Anders et al., 2014) to produce total read counts for each of the 7,126 gene features from the Ensembl annotations for each of the 7 technical replicates. For each technical replicate, the read-count-per-gene measurements were summed for each of the 96 biological replicates. Comparisons of these measurements were determined which replicates did were aberrant and thus removed from further analysis. This process yielded 42 wild-type and 44 Δ*snf2* mutant clean biological replicates (Schurch et al., 2016). Reads from these FASTQ files were then used in RNA-seq alignments with STAR.

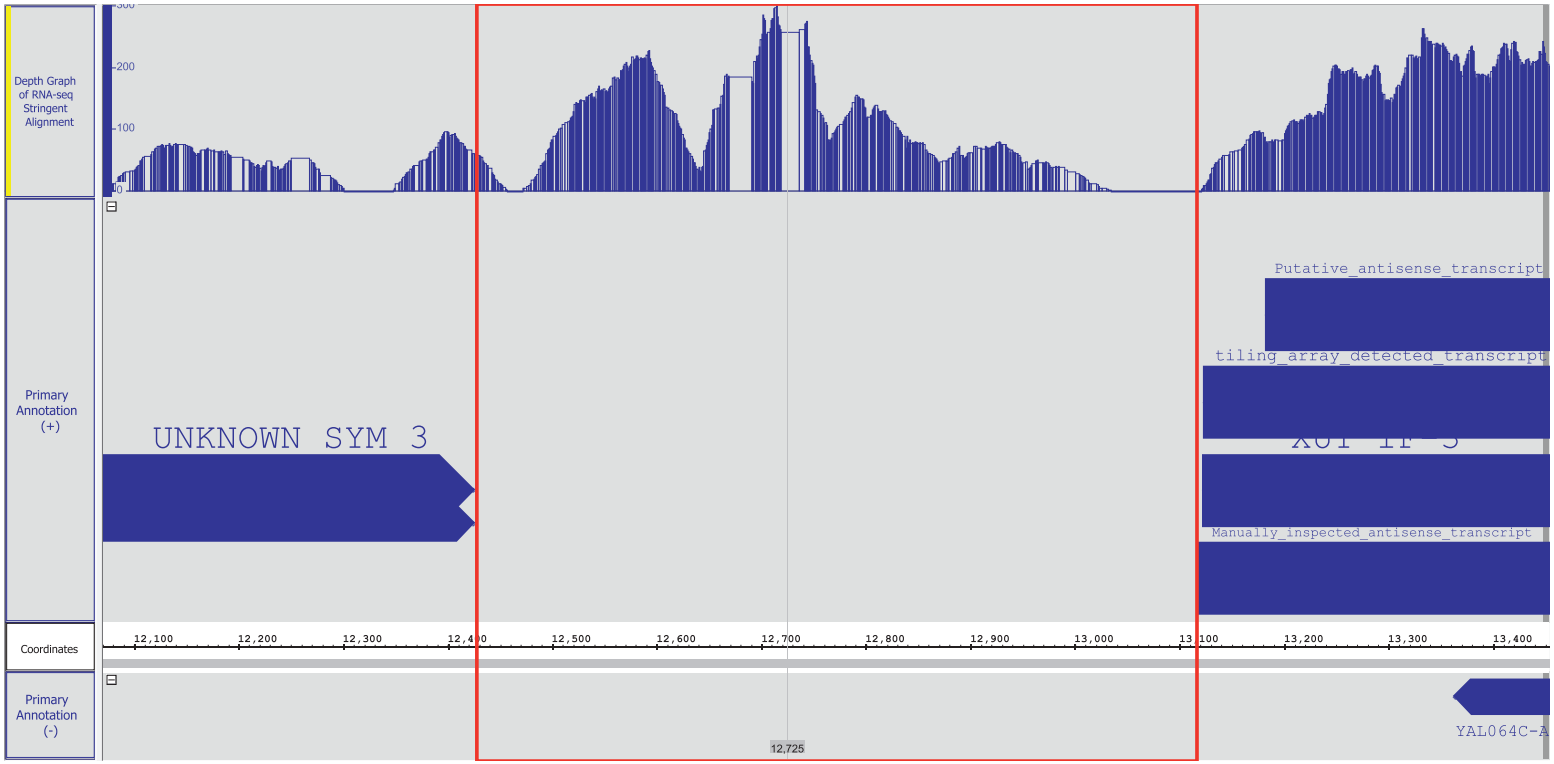
Figure 2.1: A schematic of the steps in an RNA-seq experiment.



2.2 UARs and RNA-seq Read Alignments

Regions of the genome that are highly expressed but currently un-annotated suggest the presence of new biologically relevant features in this extensively studies model species. Figure 2.2 shows an example of a highly expressed UAR on chromosome I. For some genomic positions within this region the read pileup reaches a depth of greater than 300 reads and is comparable to the expression of flanking annotated regions.

Figure 2.2: Region 12,070-13,453 on chromosome I is shown (Nicol et al., 2009). All subsequent figures showing similar information were created also in the Integrated Genome Browser. Genomic coordinates are displayed near the bottom, with Primary Annotations (see 2.4.1) immediately above (forward (+) strand) and below (backward (—) strand). Read counts of RNA-seq reads are shown above the Primary Annotation for the forward strand. The un-annotated region of interest is within the red box at 12,444-13,054, flanked by annotations upstream and downstream on the forward strand. The forward strand annotations from 5' to 3' and top to bottom are as follows: YAL064W-B (fungal-specific protein of unknown function); Putative Antisense Transcript; Unannotated Tiling Array Detected Transcript; Xrn1-Sensitive Unstable Transcript 1F-3; Manually Inspected Antisense Transcript. The reverse strand contains the YAL064C-A (putative protein of unknown function) annotation.



These potential new genomic features may include various non-coding RNAs which are functional RNA entities, such as rRNAs, snRNAs, snoRNAs, or miRNAs (Tisseur et al., 2011). Alternatively, the new features may be coding entities, such as peptides or proteins, since open reading frames (ORFs) are present in UARs.

Therefore, the aim of this study was to characterise UARs with high RNA-seq read count in *Saccharomyces cerevisiae* to determine whether they contain new genomic features. The objectives by which the study was carried out included

- process and analyse the RNA-seq data
- examine the UARs with respect to the RNA-seq data
- determine whether UARs with high levels of transcription have similar characteristics with known types of molecules

2.3 RNA-seq Data Processing

After each replicate's set of raw RNA-seq reads passes quality assessments, the reads were aligned to the reference genome to determine the transcriptional origin for each read (Wang et al., 2009). Typically, a short-read alignment program maps the RNA-seq read to the genomic location that matches best in sequence. This process is complicated by many factors, including the choice of alignment program, sequence repetition in certain regions of the yeast genome, introns, and potential sequencing errors as mentioned in Chapter 1.

2.3.1 The STAR RNA-seq Alignment Program

The RNA-seq data were previously aligned with TopHat2 (Kim et al., 2013), a splice-aware aligner. However, after several instances of failed runs with errors that could not be fixed, STAR (Dobin et al., 2013) was exclusively used from that point onward. In practice, STAR proved to run faster and have more flexibility in parameter settings. For example, TopHat2 allows for the specification of 75 different options, yet STAR allows for 108. In addition, the author of STAR was prompt in providing technical support and help with defining appropriate parameter values for the yeast genome.

2.3.2 RNA-seq Read Alignment Methodology

The 42 clean wild-type RNA-seq replicates were concatenated into a combined gzipped FASTQ file, and a STAR genomeDir directory was created to designate the names and lengths of all chromosomes for the *Saccharomyces cerevisiae* strain S288C R64 genome assembly (Mortimer and Johnston, 1986). The precise parameters chosen for an alignment can have a non-trivial effect on the final read depth distribution across the genome. To examine the effects of alignment methodology, three strategies of increasing stringency were considered. Specific values of each parameter for the alignments are found in Table 2.1.

However, there are twelve parameter values that were common to all three alignments:

- --runThreadN = 8 : This parameter sets the number of threads each alignment can run on.

- --genomeLoad = LoadAndRemove : This parameter allows the loading of the genome into the shared memory, but then removes it after the alignment is finished.
- --outSAMmode = Full : This parameter is used to set what SAM output will include. For the Full option, all SAM fields will be in the output. Choosing None would give no SAM output, and choosing NoQS would provide SAM output without quality scores.
- --readFilesCommand = zcat : This parameter indicates which type of decompression program should be used.
- --outFilterType = BySJout : This parameter has two possible values: Normal and BySJout. Normal would allow for standard output filtering using only the current alignment. BySJout keeps only the reads that passed filtering and contain junctions in the SJ.out.tab file, separating reads that align to junctions from other read alignments, reducing the number of spurious junctions detected.
- --outSJfilterIntronMaxVsReadN = 500 1000 2000 2500 : Any number of integers greater than zero may be set for this parameter. The first integer listed allows for a gap of maximum length 500 bp to be supported by only 1 read. The second integer allows for a gap of maximum length of 1,000 bp to be supported by only 2 reads, and so on. If lower values were chosen, then smaller junctions with low read counts would be allowed, and vice versa.
- --alignIntronMin = 10 : This parameter sets the minimum length of any intron - in this case, it would be 10 bp.
- --alignIntronMax = 3000 : This parameter sets the maximum length of any

intron, and in this case it would be 3000 bp.

Near-Default

Firstly, the Near-Default mapping provided a baseline of RNA-seq alignment results was prepared from a typical set of parameters appropriate for the RNA-seq dataset and the yeast genome. The following parameters were modified:

- --outFilterMismatchNoverLmax = 0.04 : This parameter is the proportion of the bp length of the read that is allowed to mismatch with the reference. For a 50-bp read this corresponds to 2-bp. If the value 0.04 were increased, more mismatches per read would be allowed, increasing the flexibility of read mapping, most likely increasing the number of read alignments in the results. If the value of 0.04 were decreased, then fewer mismatches would be allowed per read alignment, most likely decreasing the number of resulting alignments.
- --outFilterMultimapNmax = 2 : This parameter allows a maximum of 2 loci the read can align to. If there are multiple locations that a read can map to, all of these read alignments will be in the output. Therefore, if there were a single read mapping to two locations, that read would essentially be duplicated in the final alignment output. Increasing this value, for example, to 3 would allow a maximum of 3 loci any single RNA-seq read can map to, would could potentially triple the number of times a read could appear in the alignment. Reducing the value to 1 would mean that each read could map to a maximum of 1 location, and therefore, appear only once in the alignment.

Unique

If a read can map to, for example, two different locations equally well, then it is difficult to identify which which of the two regions the original RNA transcript was derived from. To eliminate this level of ambiguity, reads that could map to more than one location equally well would be excluded in the Unique alignment by setting the following parameter:

- --outFilterMultimapNmax = 1 : Instead of 2, this parameter is set to 1, now allowing for only a maximum of 1 location that each read can map to.

Stringent

Because the allowed maximum of two mismatches for each 50-bp read still allowed imperfect alignments of reads, a third set of alignment parameters was invoked to allow only reads with all bases mapping to the genome to map by setting the following parameters:

- --outFilterMismatchNoverLmax = 0 : In the Near-Default and Unique mappings, this parameter was set to 0.04, allowing a maximum of 2 mismatches per 50-bp read. However, since only a perfect match is allowed, this value must be set to 0, allowing no mismatches for any read mapping.
- --outFilterMultimapNmax = 1 : See Unique mapping.
- --outFilterMatchNmin = 51 : In addition to setting the number of mismatches allowed to 0, gaps were disallowed by setting this parameter to 51. This parameter stipulates that if and only if the number of bases matching is equal or greater than the value, the read alignment will be in the output. Therefore,

since we have 50-bp reads (which the program reads as having a length of 51 bp), each bp in the read must match in order for the alignment to be passed to the output. Setting a lower number would allow reads that have fewer matching bp to be allowed to map. Setting a number higher than 51 would allow none of the 50-bp reads to be aligned since there are a maximum of 50 bp to match.

Table 2.1: Values of relevant parameters for the STAR RNA-seq alignment program and their parameters for each of the three alignment methods are shown.

STAR Option	**Near-Default**	**Unique**	**Stringent**
--outFilterType	BySJout	BySJout	BySJout
--outFilterMultimapNmax	2	1	1
--outSJfilterIntronMaxVsReadN	500 1000 2000 2500	500 1000 2000 2500	500 1000 2000 2500
--alignIntronMin	10	10	10
--alignIntronMax	3000	3000	3000
--outFilterMismatchNoverLmax	0.04	0.04	0
--outFilterMatchNmin	–	–	51

Table 2.2: Results from STAR RNA-seq Near-Default, Unique, and Stringent Alignments for the 42 clean wild-type replicates. Percentages of the 431,650,168 total input RNA-seq reads are listed under each category. For the Stringent Alignment, 261,246,485 reads remained after further filtering for the "51M" CIGAR string. For clarification, the "Unmapped: Too Short" category refers to reads that did not reach the set minimum number of aligned bases.

Alignment	**Near-Default**	**Unique**	**Stringent**
Uniquely Mapped	84.72%	84.70%	61.17%
Mapped to Multiple Loci	10.48%	0.00%	0.00%
Mapped to Too Many Loci	2.32%	12.82%	9.17%
Unmapped: Too Short	2.44%	2.44%	29.63%
Unmapped: Other	0.04%	0.04%	0.04%

2.4 Un-Annotated Regions

In order to determine the locations of UARs, it was necessary to consider all currently known annotations for the yeast genome. Details of how the un-annotated regions were defined and the use of the Un-Annotated Region Pipeline to integrate RNA-seq analysis with the current annotations are provided in this section. In addition, some general characteristics of UARs and developed methods for sorting and categorising UARs are described.

2.4.1 Determination of the Locations of Un-Annotated Regions

The *Saccharomyces* Genome Database (SGD) houses a comprehensive set of high-quality annotations curated from the literature under Published Datasets in the Downloads section (Cherry et al., 2012). For each publication, .gff3 files were collected where available. All annotations from SGD were categorised under either Primary Annotation or Secondary Annotation.

Primary Annotation

Primary Annotations are regions of the yeast genome that are known to produce a molecular product, such as a protein or RNA molecule, or un-translated parts of transcripts (UTRs). Included under this category are also previously detected protein-coding ORFs. Table A in Appendix A provides the extensive list of Primary Annotations.

Secondary Annotation

Secondary Annotations contain information regarding potential interactions that occur at specific locations in the genome. These annotations are not characterised as previously detected transcripts or known to produce molecular entities. Interactions include meiotic recombination or crossover events and gene conversions. Binding sites of histones, transcription factors, and other proteins are also considered. Also included are modification and tagging sites, for example, polyadenylation sites and serial analysis of gene expression tagging sites, respectively. In addition, double strand break hotspots and other sequence features, such as autonomously replicating sequences and transcription start sites are included. All Secondary Annotations are listed in Table A.2 in Appendix A.

Conceivably, there could be new genomic features within regions with Secondary Annotations (e.g. a new protein-coding gene in a region with a DNA damage hotspot). Thus, the exclusion of regions with Secondary Annotations in consideration would eliminate some potential regions of interest. Therefore, Primary Annotations were used as demarcations for locations of un-annotated regions, and Secondary Annotations were treated as supplementary information.

2.5 UAR-Pipeline

To find where all UARs were located and what their characteristics were, such as length and read counts for all three alignments, the Un-Annotated Region Pipeline (UAR-Pipeline) program was written in the Python programming language to facilitate the processing and integration of multiple sources and types of data. The following Python packages were invoked in the Pipeline:

- from the Python Standard Library:
  - cPickle (serialisation of objects)
  - csv (read and write comma-delimited text files)
  - os (list files in directories)
- numpy (van der Walt et al., 2011) (numerical array operations)
- pysam (Li et al., 2009) (wrapper for samtools to read .sam and .bam files)

Workflow

Figure 2.3 shows the Pipeline's general work-flow. As input, a .gff3-, .gtf-, or .tab-formatted file with the coordinates (chromosome, start, and stop) genome annotations should be provided. The *get-uar-gff3* module then creates a new .gff3 file with the locations of all un-annotated regions.

The Pipeline can also work in conjunction with RNA-seq data after the raw reads have been mapped to the reference genome in a .bam alignment file. The module *sam-profile-uars* uses genome annotations along with the RNA-seq alignment to create a Python dictionary containing read counts per base for each un-annotated region. For better performance, the dictionary was serialised in a .pkl file with

the Python cPickle package. In Figure 2.3, this step was performed twice (once for alignment A and once for alignment B), which resulted in two separate Python dictionaries. Executing the *uar-plotly-table* module on Python dictionaries from both alignments A and B creates a tab-delimited table of all UARs. For each UAR, the table contains the chromosome, start, end, length, and total read counts over the entire region for alignments A and B.

Conversely, UAR-Pipeline can also provide read counts for specified annotated features (not described in Figure 2.3). After the Pipeline was developed, any set of annotations with any RNA-seq alignment was processed quickly and easily.

Figure 2.3: A schematic flow diagram of how the UAR-Pipeline works. Files are enclosed in boxes with their respective formats in brackets, and functions used to produce the subsequent outputs are in italics (see Table B.1).

2.5.1 Pipeline Modularization and Command-Line Usage

Several Python functions were written in a main script called process_annotation.py. Lower-level functions were written first, for example, to parse genome annotations and determine locations of UARs in the genome. By writing the lower-level functions separately, they were easily used in many larger functions, for instance, to count reads in RNA-seq alignment files for the UARs.

Furthermore, a wrapper script called uar_pipeline.py was written to turn process_annotation.py into a command-line application, the UAR-Pipeline itself, for use in the Linux/Unix environment. The underlying structure of the process_annotation.py script facilitated the modularisation of the Pipeline. In other words, the Pipeline contains 16 different programs (or modules) that execute the larger functions within the main script (Table B.1 in Appendix B). Throughout the study, new modules, often variations of previously existing ones, were written in process_annotation.py and subsequently added to the uar_pipeline.py wrapper without difficulty.

The Python packages called in the wrapper are sys (from the Python Standard Library) to append directories to the current working directory and the standard_parser and standard_logger modules written by Dr. Nick Schurch. The latter two modules were written to streamline the usage of the argparse, warnings, tempfile, and logging Python Standard Libraries together.

Moreover, a few programs were written in another script, graphical_analysis.py, to create plots of the data, for example, the *uar-lengths-hist* program. This script also called the Python math Standard Library and the matplotlib package (Hunter, 2007). Data visualisation was also undertaken by producing output to a .csv or .tab

file and then plotting with the ggplot2 package (Wickham, 2009) in R.

2.5.2 Unit Testing

In order to ensure functions in the process_annotation.py script were yielding correct output, twelve unit tests were written in the script test_process_annotation.py with the unittest Python Standard Library. Artificial inputs were created for many of the tests, and assert statements were employed to test whether the actual outputs from the functions matched the expected outputs. For example, a simplified .gff3 annotation file was created and treated as input for functions that find UARs. The function's output UAR coordinates were then compared to the known un-annotated regions.

These unit tests were crucial throughout the study when pre-existing functions were modified for better performance or technical issues, such as the need to use a more updated version of Python, which necessitated the upgrade of the version of pysam. The newer version of pysam left some previous functions used in UAR-Pipeline scripts defunct, requiring a modification of how .gff3 and .gtf files were parsed.

Successful unit testing would also allow for future users to add to and modify the current scripts without changing the correct output.

2.6 SGD Features and UARs

2.6.1 Characteristics of SGD Features

The term SGD Features refer to SGD protein-coding genes, rRNA, snoRNA, and snRNA. In this section, characteristics of each SGD Feature type are examined and compared against those of Un-Annotated Regions.

SGD Protein-Coding Genes

On a chromosome level, the lengths of SGD genes are represented as boxplots in Figure 2.4. Chromosome X has the shortest minimum length, whereas the mitochondrial chromosome has the longest. Chromosome XII had the longest SGD gene, whereas chromosome I had the shortest maximum length. Strikingly, the mitochondrial chromosome has the highest mean length at 2573 bp. The overall minimum, maximum, mean, and median lengths of all SGD genes across all chromosomes are 51 bp, 14,733 bp, 1,764 bp, and 1,071 bp, respectively. The number of SGD genes per 100 kbp on each chromosome is shown in Figure 2.5. Almost all chromosomes had between 50-60 SGD genes per 100 kbp, except the mitochondrial chromosome at about 30.

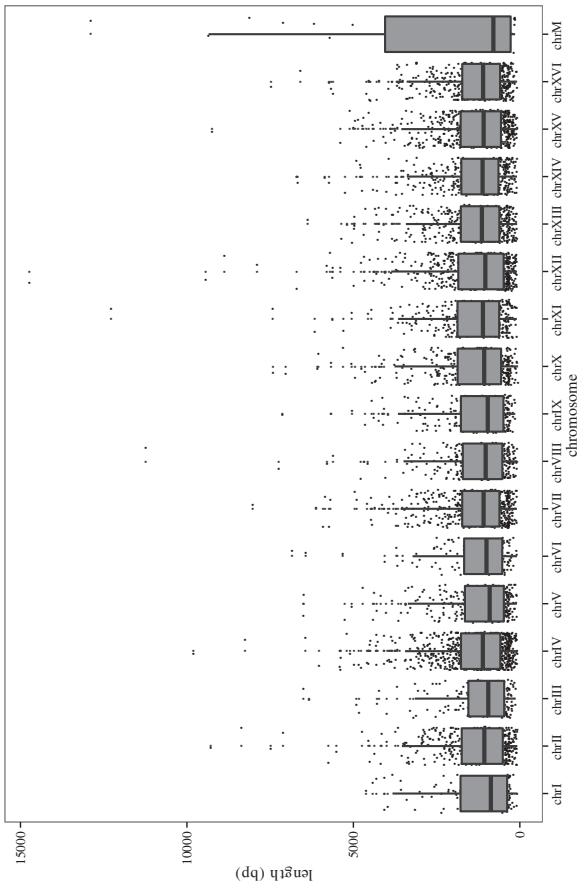


Figure 2.4: Boxplots of lengths of SGD potein-coding genes per chromosome.

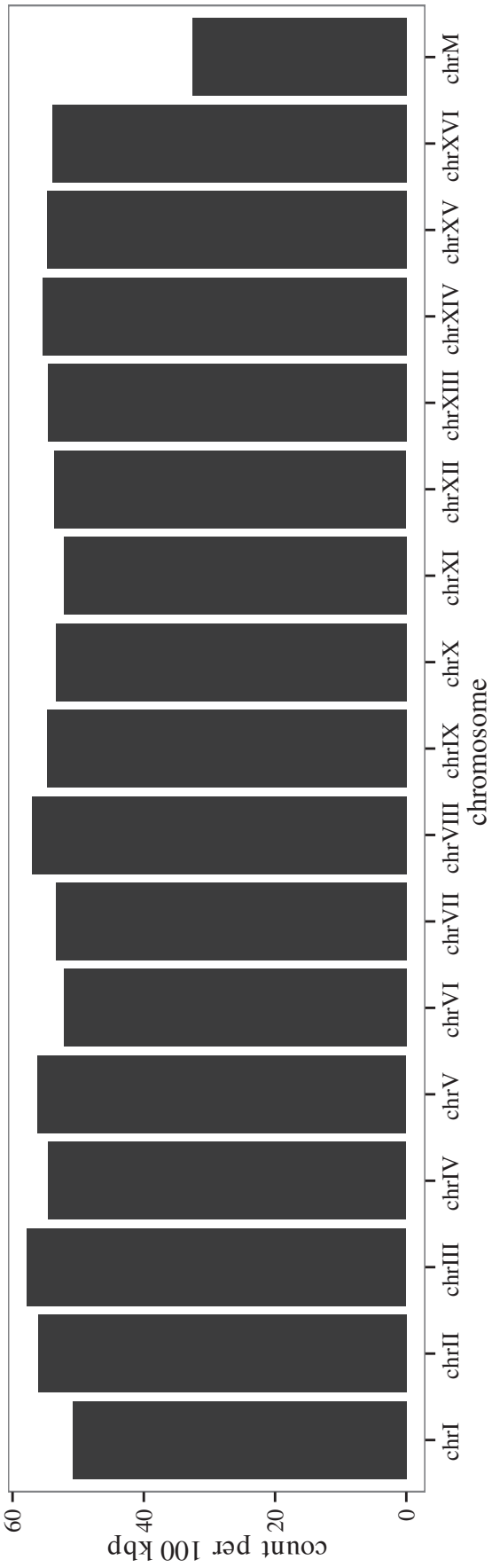


Figure 2.5: Number of SGD protein-coding genes per 100 kbp, calculated per chromosome.

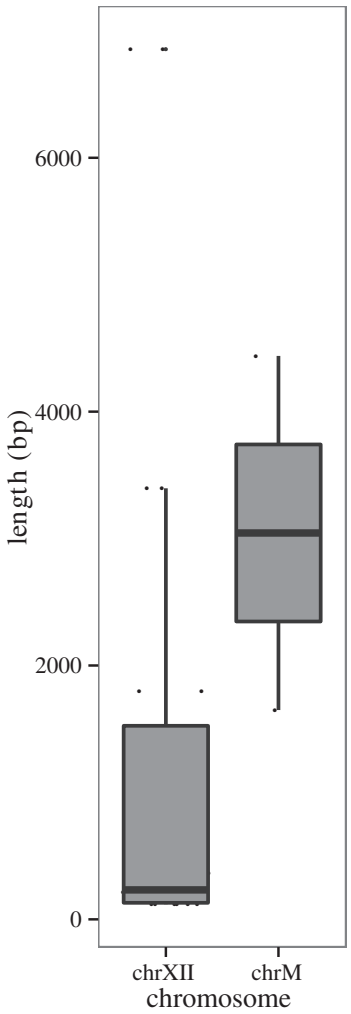
SGD rRNA

As rRNAs occur only on chromosomes XII and M (mitochondrial), boxplots to illustrate the distributions of the lengths of rRNAs for these two chromosomes are shown in Figure 2.6. The median length of ribosomal RNAs on chromosome M are longer; however, the longest rRNAs overall occur on chromosome XII. There are between 2-2.5 rRNAs per 100 kbp on chromosomes XII and M.

SGD snoRNA

Boxplots for the lengths of snoRNAs across each chromosome are plotted in Figure 2.7. Median lengths for each chromosome occur between 100-200 bp. The longest snoRNA occurs on chromosome XIII at 1,004 bp. There is a more variable density of snoRNAs across chromosomes 2.8, with chromosomes II, IV, and IX having the lowest densities and chromosome chromosome XIII with the highest density.

Figure 2.6: Distributions of the lengths of SGD rRNA.



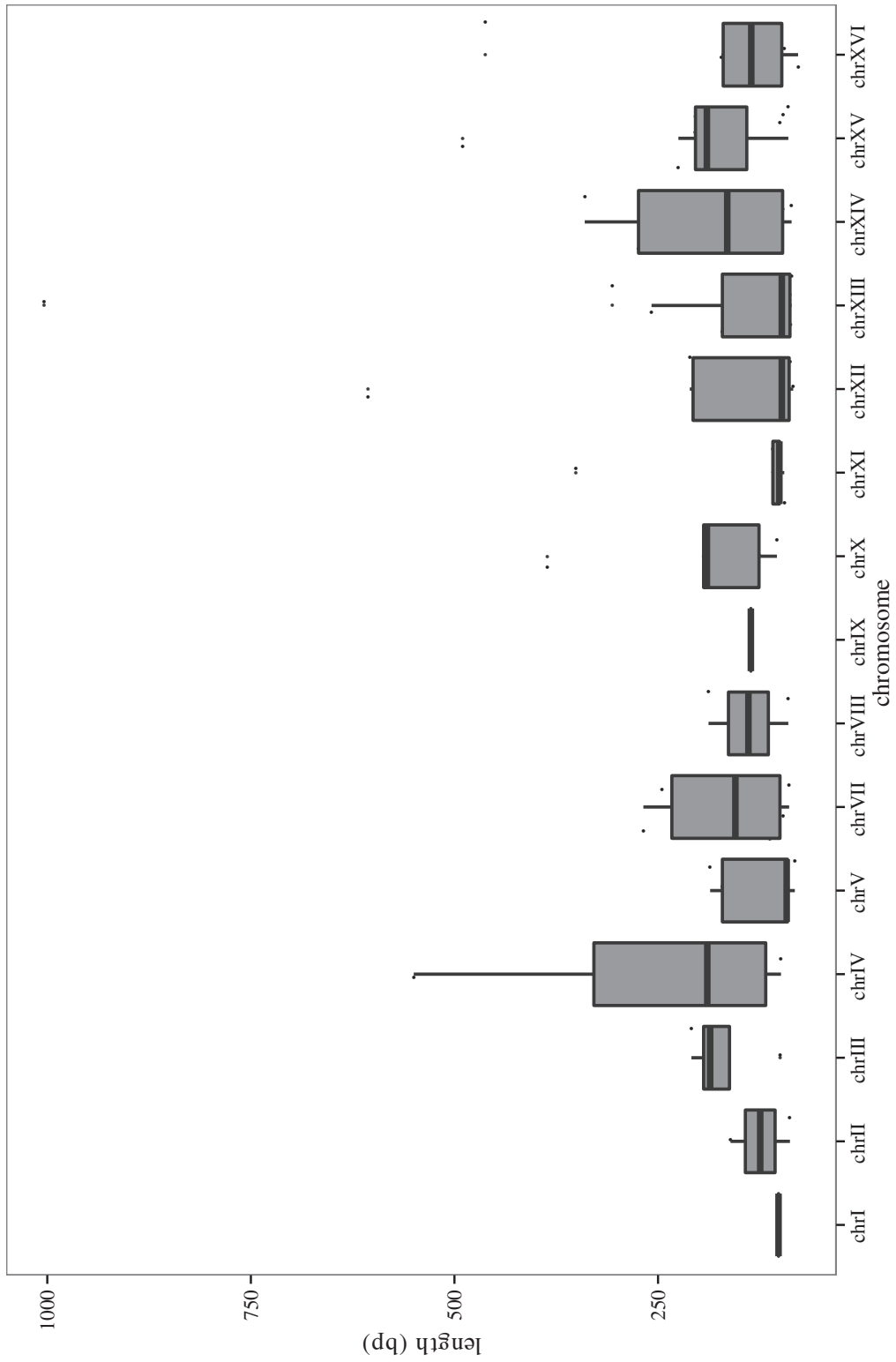


Figure 2.7: Distributions of the lengths of SGD snoRNAs.

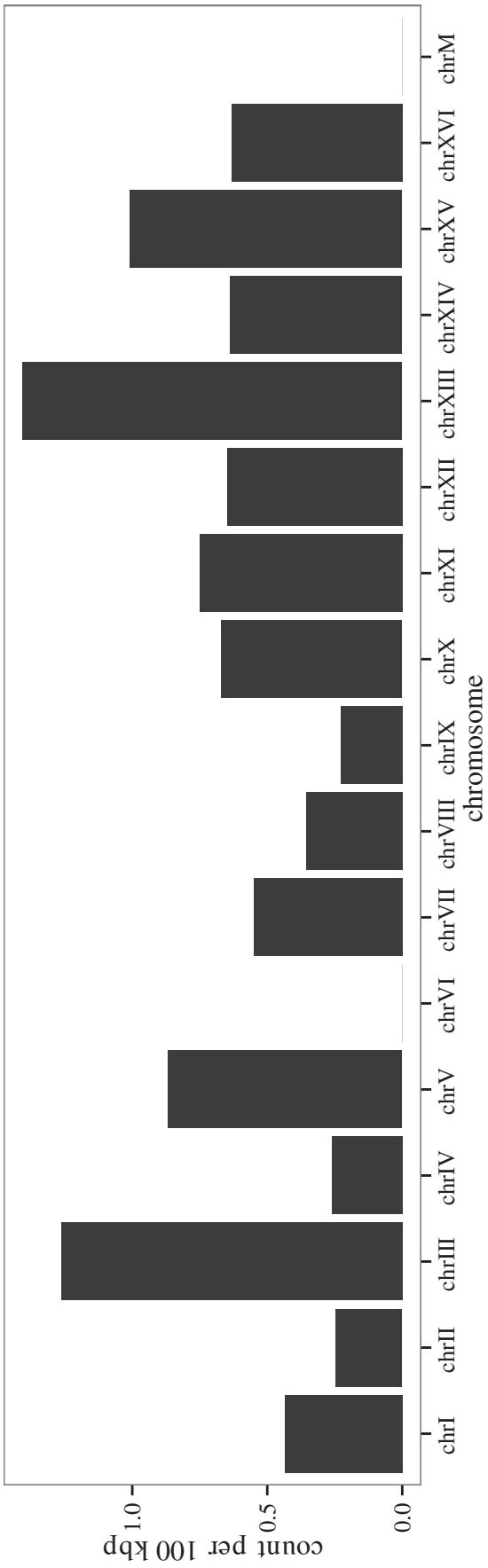
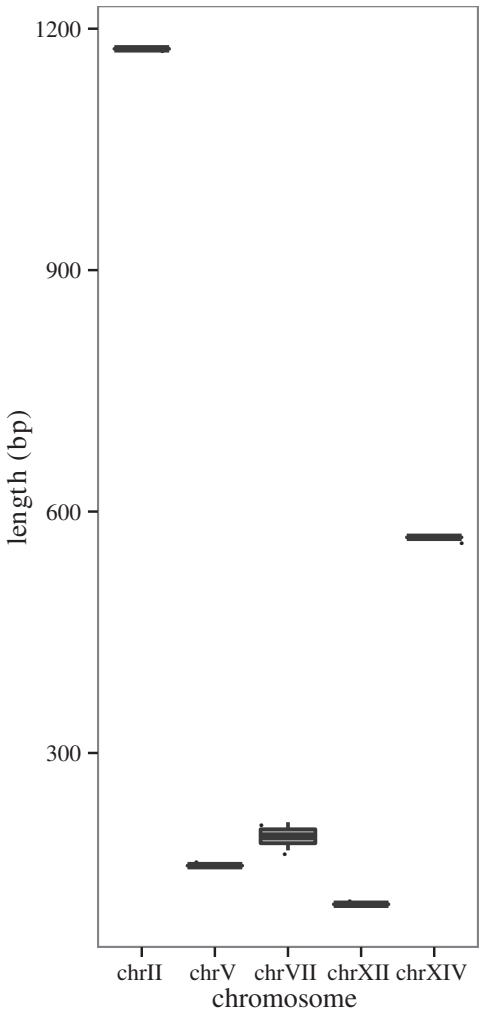


Figure 2.8: Number of SGD snoRNA per 100 kbp, calculated per chromosome.

SGD snRNA

The snRNAs occur only on chromosomes II, V, VII, XII, and XIV, with each chromosome having only one snRNA, except chromosome VII, which has two. Of these chromosomes, chromosome II has the longest snRNA molecule at 1,175 bp, whereas chr XII has the shortest at 112 bp 2.9. As there are only 6 snRNAs across the entire genome, it is expected that the densities of snRNAs for each chromosome would be very low, as seen in 2.10.

Figure 2.9: Distributions of the lengths of SGD snNAs.



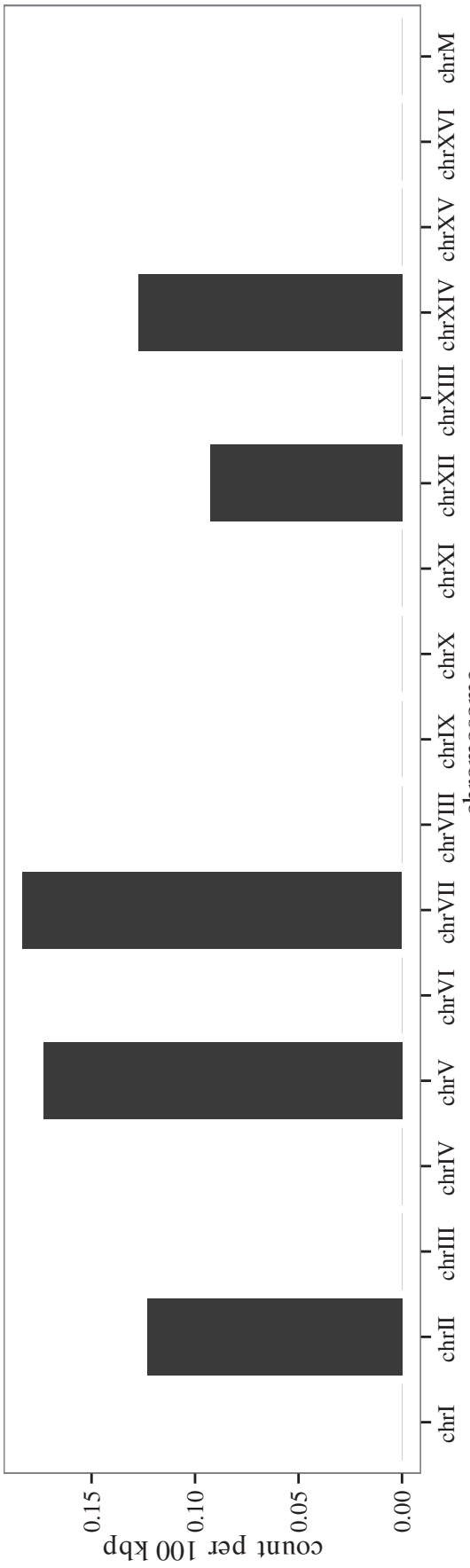


Figure 2.10: Number of SGD snRNA per 100 kbp, calculated per chromosome.

2.6.2 Un-Annotated Regions

On a per-chromosome basis, the lengths of all UARs were plotted in Figure 2.11. Chromosomes I and M have much larger spreads of lengths than do the other chromosomes, with chromosome XV having the longest UARs at 2,160 bp. Interestingly, chromosome M has the highest density of about 60 UARs per 100 kbp (Figure 2.12), whereas all other chromosomes have around 20-30.

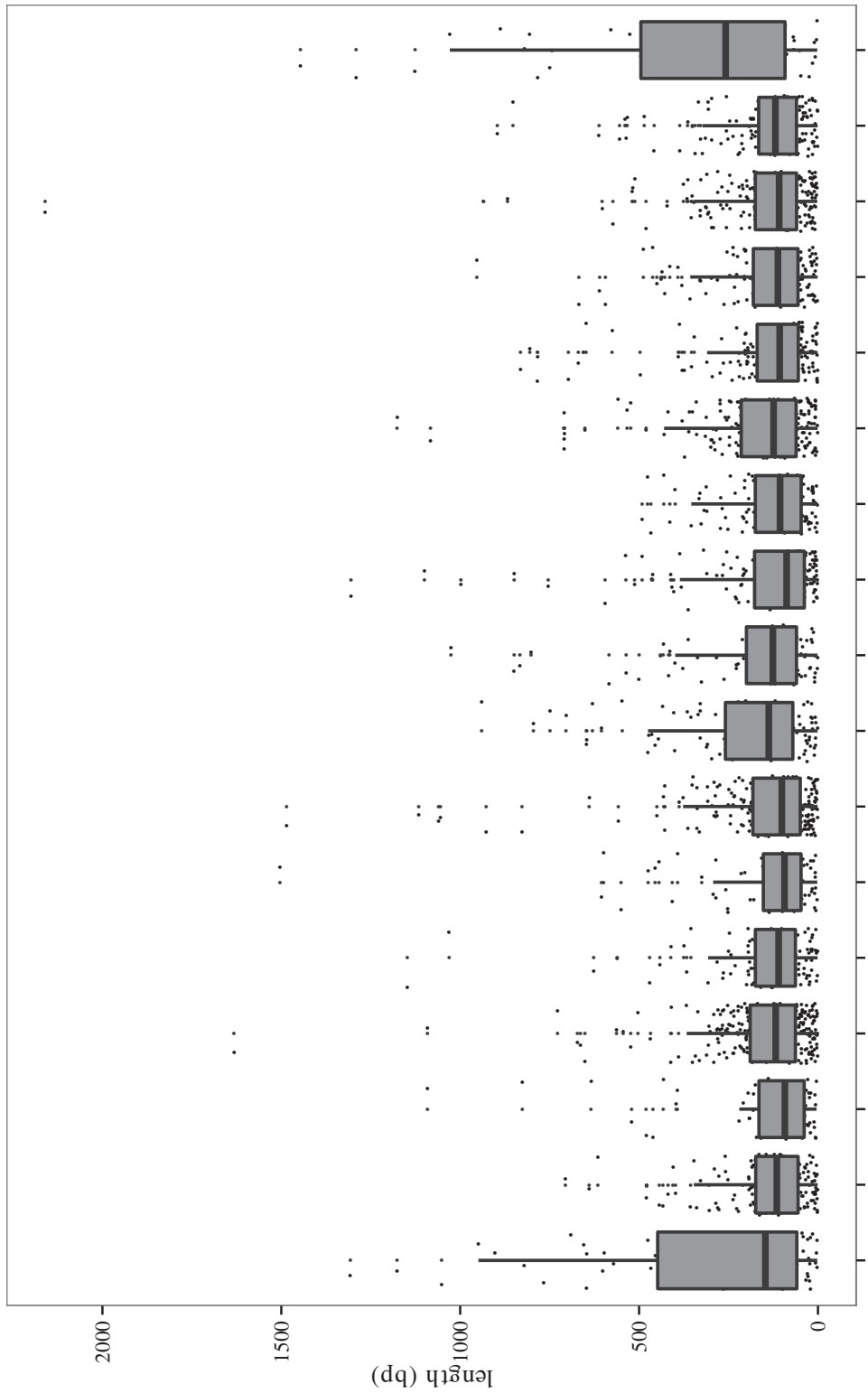


Figure 2.11: Distributions of the lengths of un-annotated regions.

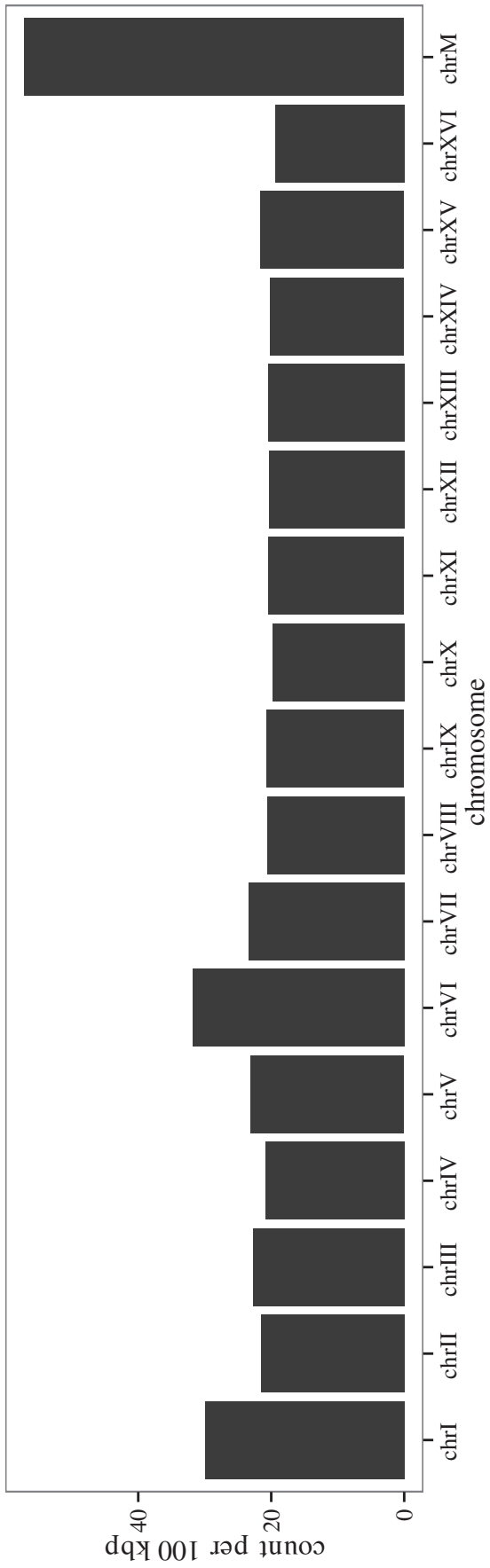


Figure 2.12: Number of un-annotated regions per 100 kbp, calculated per chromosome.

2.6.3 Comparisons of SGD Features and Un-Annotated Regions

Figure 2.13 allows the comparisons of the distributions of lengths of SGD Features and UARs, and UAR ORFs (open reading frames derived from UARs). Overall, UAR ORFs have distributions centered on the shortest lengths, followed by snoRNA/snRNA/UARs, rRNAs, and lastly by SGD genes which have the highest concentration of the longest lengths. For a clearer comparison, all RNA SGD Features were removed to show the distributions of SGD protein-coding genes, UARs, and UAR ORFs in finer detail in Figure 2.14. On all chromosomes, UAR ORFs have the shortest lengths, followed by UARs, and then SGD genes, as mentioned previously.

Figure 2.13: Probability distributions of the lengths of SGD protein-coding genes ('gene'), rRNA, snoRNA, snRNA, un-annotated regions ('UAR'), and un-annotated region open reading frames ('UAR_ORF').

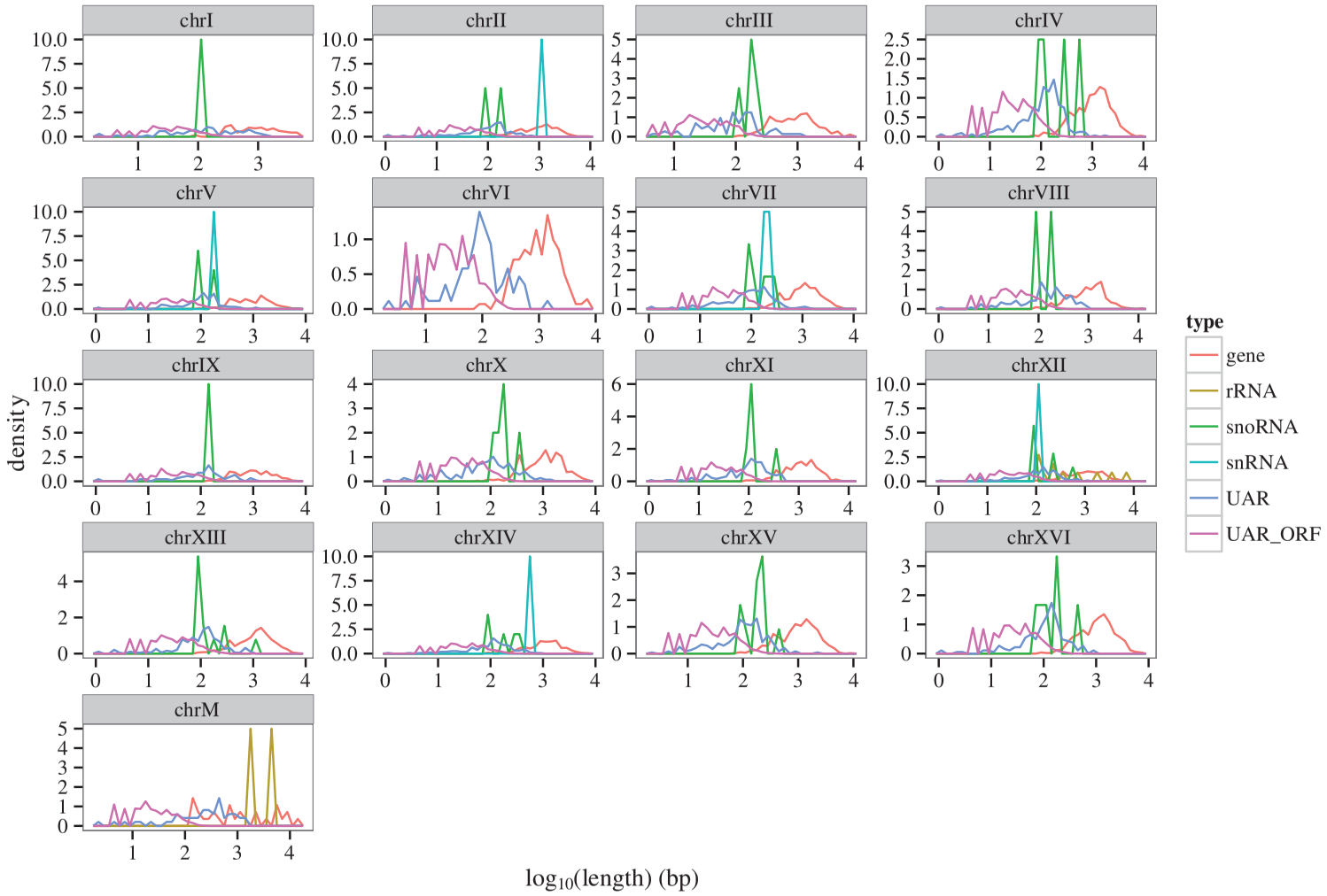
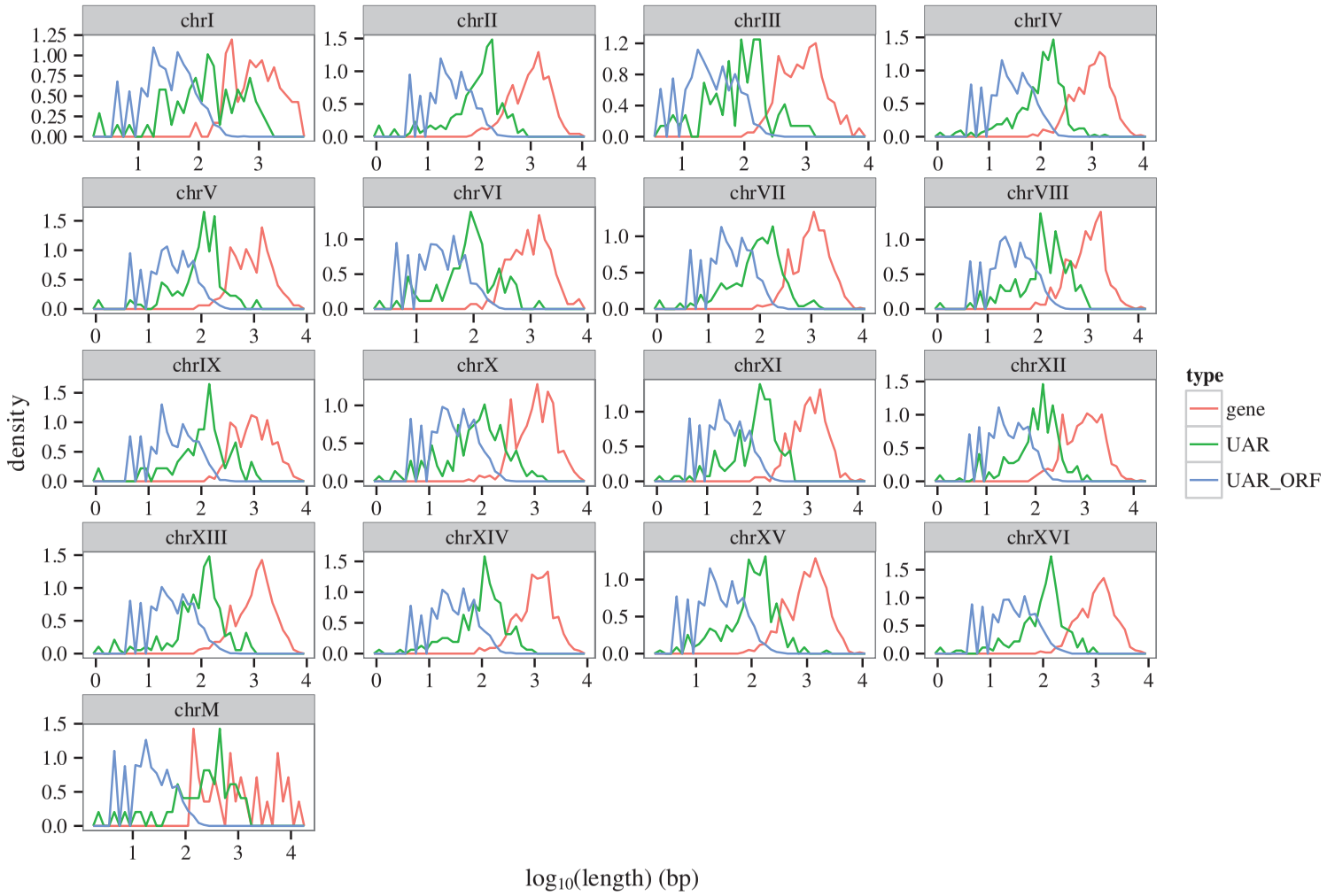


Figure 2.14: Probability distributions of the lengths of SGD protein-coding genes ('gene'), un-annotated regions ('UAR'), and un-annotated region open reading frames ('UAR_ORF').



2.6.4 Comparison of the Three Methods of RNA-seq Mapping

The distributions of read counts for SGD protein-coding genes for the Near-Default, Unique, and Stringent mapping methods per chromosome were analysed (data not shown). For almost all chromosomes except the mitochondrial chromosome, there is great overlap amongst the distributions for the three mappings. In chromosome M, the Stringent distribution is shifted to the left (lower read counts) as compared to the Near-Default and Unique distributions. Distributions for rRNAs on chromosomes XII and M are plotted in Figure 2.15, where there are also shifts toward lower read counts for Unique and Stringent mappings in reference to the Near-Default distributions. These trends are expected since the Unique and Stringent mappings have stricter criteria for read mapping, so successively fewer reads will remain in these alignments.

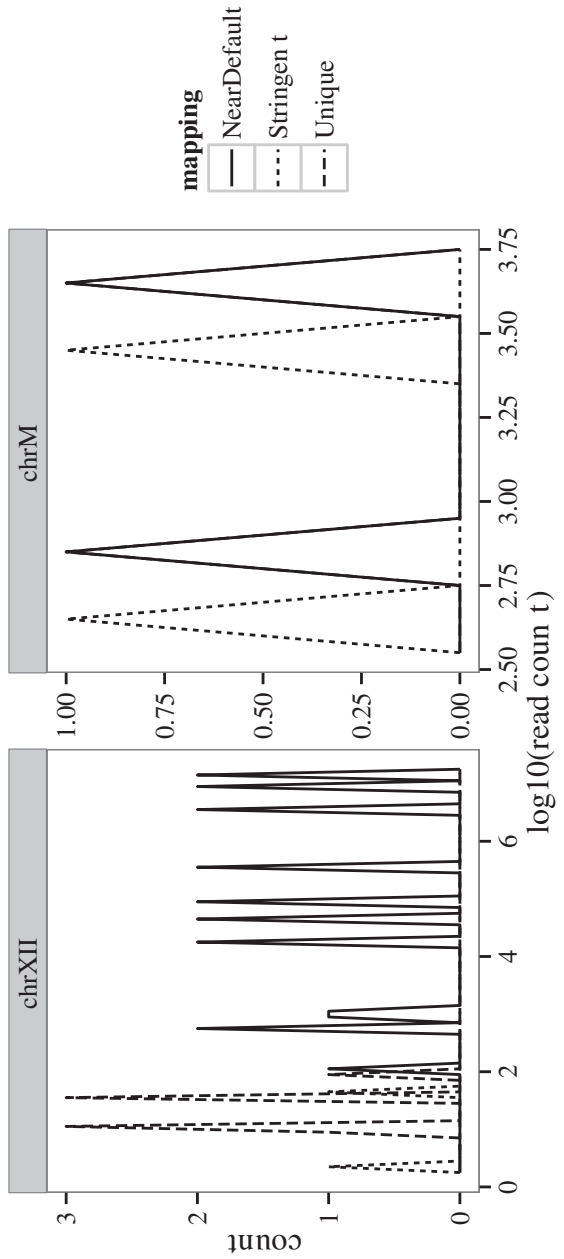
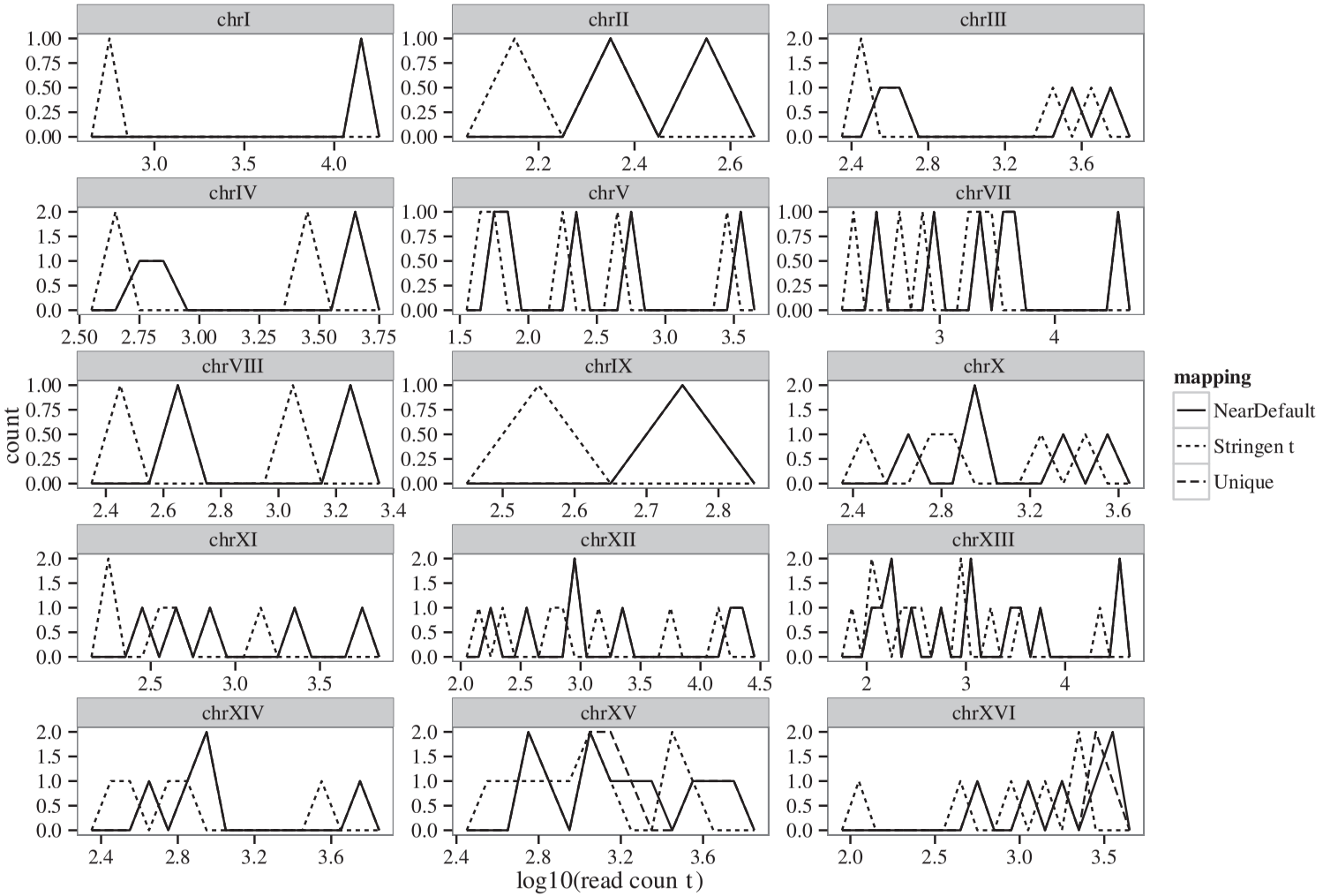


Figure 2.15: Distributions of read counts for rRNAs on chromosomes XII and M.

Similar trends are evident for snoRNAs (Figure 2.16), with chromosome XVI having the lowest read counts and chromosome XII having the snoRNAs with the highest read counts overall. For completeness, these distributions were plotted for snRNAs as well in Figure 2.17.

Figure 2.16: Distributions of read counts for SGD snoRNA genes for the Near-Default, Unique, and Stringent mapping methods.



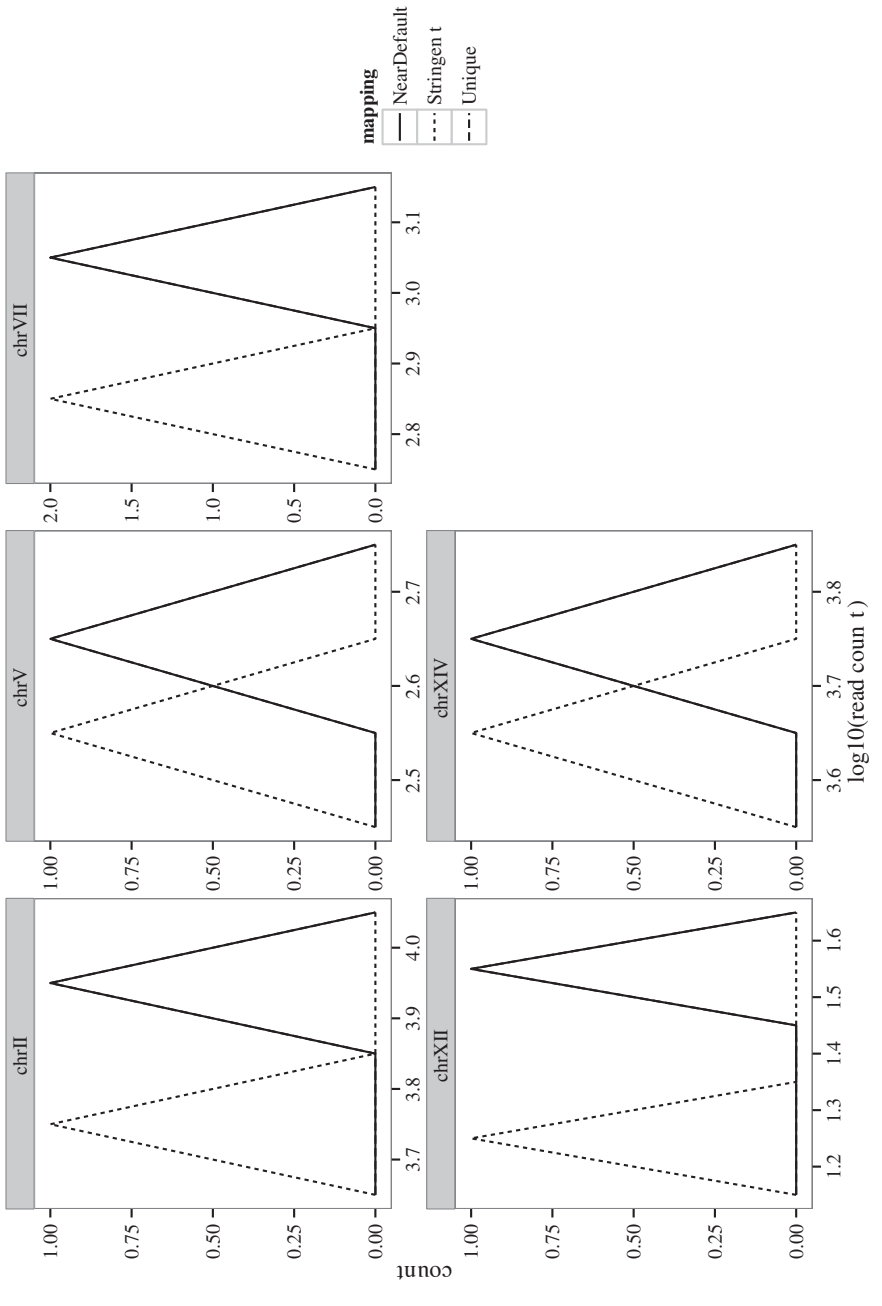


Figure 2.17: Distributions of read counts for SGD snRNA genes for the Near-Default, Unique, and Stringent mapping methods.

Further comparisons of SGD Features and Un-Annotated Regions can be made with Figure 2.18, in which scatter plots of Near-Default Read Counts vs. Length are shown. The primary trends illustrated in this panel are that SGD Features generally are at least 100 bp in length with over 100 read counts. However, for UARs, there is a wider spread of lengths and read counts, with the majority of the distribution concentrated at lower values for both axes. Similarly, a panel of graphs for the Unique mapping shows that rRNAs shifted downward toward lower read counts (Figure 2.19). This may occur since there is a lot of sequence homology within and amongst rRNA genes; thus, a single 50-bp read may map to multiple regions equally well. These reads would be excluded in the Unique mapping, causing a decrease in the overall read count per rRNA. Comparing the scatter plots for the Stringent mapping (Figure 2.20) to the Unique mapping, there do not seem to be large differences.

Figure 2.18: Scatter plots of Near-Default Mapping Read Counts against Length for un-annotated regions and SGD protein-coding genes.

Figure 2.19: Scatter plots of Unique Mapping Read Counts against Length for un-annotated regions and SGD protein-coding genes.

Figure 2.20: Scatter plots of Stringent Mapping Read Counts against Length for un-annotated regions and SGD protein-coding genes.

2.7 Analysis of Un-Annotated Regions

There are a plethora of classical methods and tools for determining properties of protein sequences, including InterPro and BLAST (Jones et al., 2014) and (Altschul et al., 1990). InterPro classifies protein sequences in families for functional analysis and predicts domains and important sites by integrating information from a multitude of databases (Section 1.6.1). BLAST invokes sequence alignment to find other sequences in databases of interest that closely match with the query sequence (Section 1.6.1). Tools such as these are heavily applied in many genomic studies including those with the aim of predicting new un-annotated genes in yeast (ÒhÈigeartaigh et al., 2011). In some instances, if newly predicted genes are not functionally verified, they become labelled as predicted, putative, uncharacterised, or unknown in current sets of annotation from SGD.

Therefore, searching UARs against current databases may yield many of these types of annotations, providing very little useful information about the query sequence. Moreover, if, for instance, one UAR is searched against InterPro and yields no result, it may be that the UAR would be the first of its type of signature to be detected and would not have any previous related information. A sorting method was developed to prioritise UARs with the best evidence of having a new genomic feature. This method included other sources of information, such as MULTIZ (Blanchette et al., 2004) and phastCons (Siepel et al., 2005), which do not rely on these tools. MULTIZ is a program that creates local alignments in a pair-wise fashion for multiple sequences (Section 1.6.1). In this case, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castelli*, and *S. kluyveri* were aligned to *S. cerevisiae* individually. The

MULTIZ score is the number of these *Saccharomyces* species that aligned at each base in the *S. cerevisiae* genome. The phastCons score is a conservation score per base based on MULTIZ7way alignments (Section 1.6.1). The conservation scores are predicted from fitting a phylogenetic hidden Markov model to the data by maximum likelihood (Section 1.6.1). Therefore, a higher score means a higher probability that the base is conserved across the 7 *Saccharomyces* species.

All 2,636 UARs were sorted successively by the following characteristics (Table 2.3):

- expression (RNA-seq read counts in the Stringent alignment)
- maximum MULTIZ score
- phastCons sum
- length of the UAR

The logic behind the specific order of sorting is that RNA-seq expression, an indication of transcription, would be the primary driving factor behind the potential existence of a new genomic feature. Secondly, the higher the number of other yeasts that could align at some point within the UAR, the higher the chances are that the UAR is conserved. Conservation would give further evidence toward a viable genomic feature, and make it more probable that the feature would also be functional. Thirdly, the sum of phastCons scores across all the bases within an UAR gives another indication of the probability of conservation. Also, the longer the UAR, the more likely it is to contain a new feature. Successive sorting has an effect on the order of UARs only if multiple regions contain the same number of RNA-seq read counts.

Table 2.3: All 971 un-annotated regions were sorted firstly by the read count in the RNA-seq Stringent alignment, secondly by the maximum MULTIZ score, thirdly by the sum of phastCons scores, and lastly by the length of the UAR. This table shows the first 20 UARs.

Chr	**Start**	**End**	**Length**	**Near-Default Read Count**	**Unique Read Count**	**Stringent Read Count**	**MULTIZ Score**	**MULTIZ Max Score**	**phastCons Score**
chrI	12427	13117	691	3922	1826	1322	2758.0	6	26.367
chrVIII	542362	543000	639	8399	1305	966	3195.0	5	307.14
chrVIII	17842	18636	795	1359	1359	962	3691.0	5	21.819
chrIII	314982	315162	181	1791	1014	753	724.0	4	32.062
chrVII	1072316	1072994	679	992	992	708	130.0	3	0.477
chrVII	404448	404785	338	1782	820	606	1395.0	5	125.086
chrXV	1058591	1059002	412	690	690	508	1270.0	4	52.254
chrIV	805141	805643	503	1289	688	497	2574.0	6	443.971
chrII	460372	460656	285	685	685	483	831.0	3	12.365
chrVI	258250	258854	605	662	662	477	1763.0	3	11.725
chrVI	259767	261013	1247	581	581	431	4921.0	4	431.094
chrI	22686	23992	1307	1019	596	414	2880.0	4	54.645
chrVII	1068995	1069994	1000	551	551	412	4744.0	5	92.276
chrXI	567359	567749	391	531	531	388	1955.0	5	34.217
chrXII	365457	365768	312	534	534	379	267.0	3	24.595
chrXII	809186	809718	533	479	479	355	2313.0	5	330.956
chrVII	1083577	1083634	58	486	483	338	116.0	2	4.302
chrIX	341685	341856	172	463	463	337	424.0	4	7.489
chrI	226923	227523	601	472	472	324	2179.0	5	518.114
chrI	20390	21565	1176	475	444	319	3781.0	4	303.68

2.8 IGB QuickLoad Site

In this study, the volume of RNA-seq data and the number and variety of genomic annotations necessitated a way to access and visualise the data quickly and easily. Viewing all the annotation data, the genomic sequence, and the high depth RNA-seq data in the context of the genome is a challenging problem. We use the open-source stand-along genome browser Integrated Genome Browser (IGB) for this purpose (Nicol et al., 2009). A particularly useful feature of this software is the ability to construct your own IGB QuickLoad server site for data collections that can then be accessed simply (and privately if necessary) from within IGB. A QuickLoad site was constructed to contain all RNA-seq alignments and genomic annotations, as well as other pieces of information such as the MULTIZ (Blanchette et al., 2004) phastCons (Siepel et al., 2005) scores. The QuickLoad site can become a resource for the research community.

The structure of this section reflects first level of organisation in the QuickLoad site.

2.8.1 Primary Annotations

Primary Annotations described in Section 2.4.1 are included in this folder. There is a combined .gff3.gz file that contains all of the Primary Annotations for a more comprehensive view.

Individual Tracks

Within the Primary Annotations folder is a subfolder containing each individual annotation track as listed in Appendix Table A. Individual tracks are convenient to use in IGB if only snoRNAs, for example, were studied at a particular instance. In addition, different features may be coloured differently for emphasis.

2.8.2 Secondary Annotations

Secondary Annotations, as described in Section 2.4.1, are contained within this folder. As there were a total of 58 individual annotation tracks, listed in Appendix Table A.2 and labelled with superscripts, they were sub-divided into seven categories to decrease search time for specific types of annotations. The seven categories are listed as and described in the following subsections. Within each category, there is also a combined .gff3.gz file, named after the category, that contains all annotations within that specific category.

DNA Damage

The DNA Damage category mainly includes annotations related to double-strand break hotspots.

DNA-DNA Interactions

DNA-DNA Interactions contains annotation tracks indicating locations of meiotic recombination hotspots, gene conversions, and meiotic crossovers.

Histone Binding Sites

Histone binding sites can be inferred from nucleosome positions and mapped mononucleosomal fragments, which are included in this category.

Modification or Tagging Sites

Polyadenylation sites and serial analysis of gene expression tagging sites are indicated in the tracks within this category.

Other Binding Sites

This category contains binding sites of proteins, other than histones, that were found by methods such as ChIP.

Other Sequence Features

This category includes the locations of autonomously replicating sequences and their consensus sequences, TATA elements, and transcription start sites.

Transcription Regulation

Transcription factor binding sites determined by ChIP-chip from the Mayer et al. (2010) study are listed here.

2.8.3 RNA-seq Alignments

Under the sub-folders GRNAseq and WT are the three RNA-seq alignments described in Section 2.3.2.

2.8.4 Protein Coding ORFs

Tracks containing all possible open reading frames in all six frames of translation obtained from the EMBOSS Transeq program (Rice et al., 2000; Goujon et al., 2010) are in this category. There are two tracks: one is labelled "relaxed," in which ambiguous codons (appearing as "X" in the Transeq output) were considered coding amino acids instead of stop codons, whereas in the other, ambiguous codons were treated as stop codons.

The purpose of these tracks is to allow the user to visualise genomic locations where potential peptides and proteins may be coded from alongside the RNA-seq data and other annotations that may indicate the presence of a transcript, such as a polyadenylation site at the 3' end of an ORF.

2.8.5 Conservation

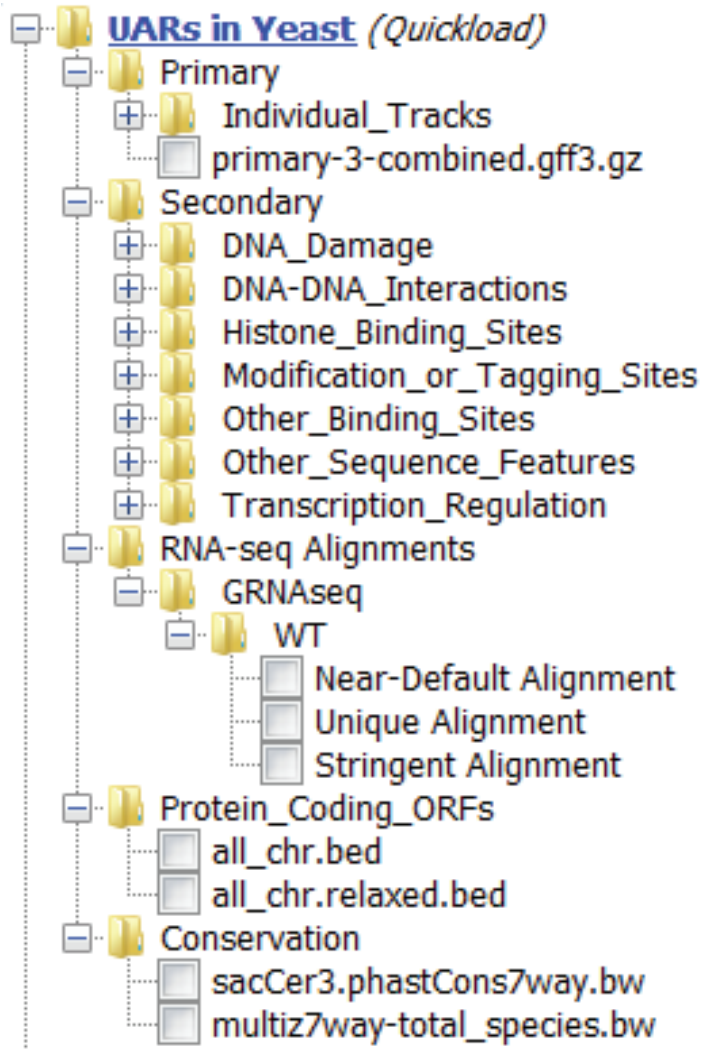
There are two .bigwig tracks in this category related to conservation. The first gives the total number of *Saccharomyces* species in the 7-way MULTIZ alignments per base, described in Section 2.7. In addition, the phastCons scores per base are also given in a separate track.

2.9 QuickLoad Site Usage

When the QuickLoad Site has been loaded into the Integrated Genome Browser, it will have the file structure shown in Figure 2.21.

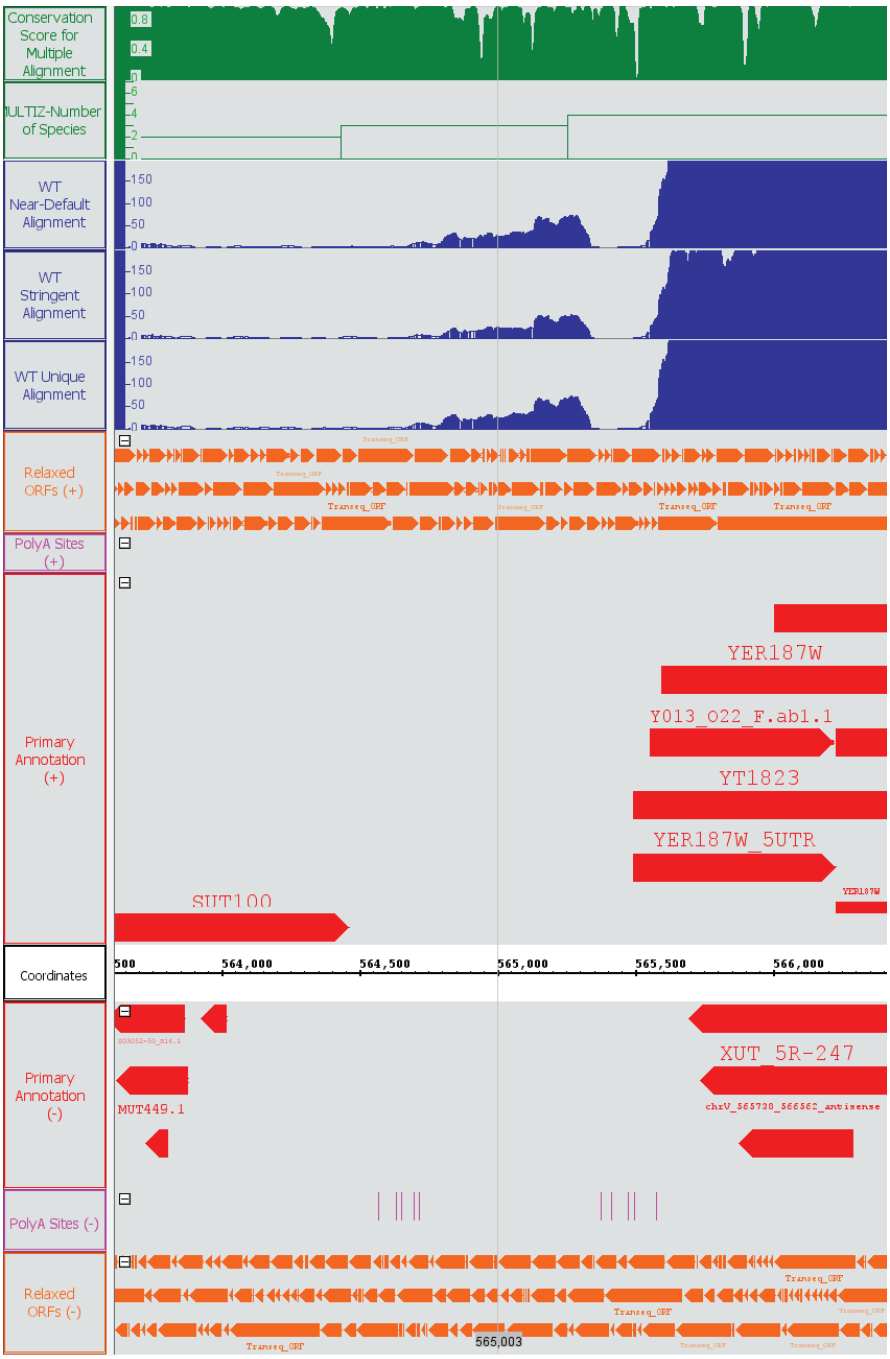
Figure 2.22 highlights the advantages in how the QuickLoad Site is structured

Figure 2.21: General file structure of the QuickLoad Site.



and the ability to access and arrange a high volume of information. The Primary Annotations indicate where the UAR is located in addition to which known annotations flank the UAR. The RNA-seq read alignments show where reads mapped to in the three stringency levels that can be compared directly. The presence of polyadenylation (polyA) sites gives evidence of the presence of a transcript(s) and where the ends of transcripts may be. All open reading frames are delineated, which indicate where potential peptides and proteins may be encoded. Lastly, the MULTIZ and phastCons tracks provide visual representations of numerical data regarding conservation at each base in the entire yeast genome. Specifically in Figure 2.22, one hypothesis maybe that the small continuous region of RNA-seq reads at around 564,700-565,400 indicates the presence of a previously undetected transcript. Given that there are polyA sites on the reverse strand, it is plausible that the new reverse strand transcript is polyadenylated at about 564,700. Interestingly, in addition to the group of polyA sites at 564,700, there is another cluster around 565,500, which most likely corresponds to the 3' ends of XUT_5R-247, chrV_565736_566562_antisense, and other annotations shown. However, as there are no other annotations downstream of these transcripts, it is also plausible that these transcripts may be incorrectly annotated and that their 3' ends may in fact be further downstream, toward the set of polyA sites around 564,700. These scenarios illustrate the power of the QuickLoad Site in genomic and RNA-seq analysis because of the vast amount of information it contains coupled with the flexibility in visualising data in the Integrated Genome Browser.

Figure 2.22: This IGB screenshot illustrates one way of arranging and displaying annotations, RNA-seq alignments, and conservation information to more effectively analyse the UAR chrV: 564463-565493. Starting from the Coordinates line, Primary Annotations are placed directly above (forward strand) and below (reverse strand). Immediately adjacent to the annotations are polyadenylation sites (Ozsolak et al., 2010), then all potential open reading frames in all six frames from the relaxed track. Since the RNA-seq data were unstranded, all three alignments were stacked on the forward strand. Lastly, the track giving the number of *Saccharomyces* species in multiple genome alignments with the MULTIZ program (Blanchette et al., 2004) and the phastCons (Siepel et al., 2005) scores are shown.



Chapter 3

Preliminary Targets

3.1 Introduction

This chapter illustrates analytical methods for finding and characterising interesting un-annotated regions using three prime examples. All three were found from a previous version of the Primary Annotations that was based on the EF4.70 Ensembl annotations (Flicek et al., 2014), un-translated regions (Nagalakshmi et al., 2008; Yassour et al., 2009), transposons (Cherry et al., 2012), and long-terminal repeats (Cherry et al., 2012). Also, the RNA-seq mappings were performed with TopHat2, prior to the switch to STAR. However, throughout the chapter, the target UARs were also characterised in the context of the current Primary Annotations and STAR alignments where necessary.

The first target UAR was found through analysing statistics (mean, median, standard deviation, and standard error) across the 42 clean WT biological replicates. The UAR chrXII: 489,949–490,404 bore similarity to sequences of rRNAs through a nucleotide BLAST search. The second UAR at chrI: 12,427–13,361 had homology

Table 3.1: The sets of annotations referred to in this section and their contents.

Annotation Set Name	**Contents**
EF470_UTRs.gtf	Ensembl version EF4.70 annotations; untranslated regions
EF470_UTRs_transposons_LTRs.gtf	Ensembl version EF4.70 annotations; untranslated regions; transposons; long terminal repeats
Primary Annotations (current)	(see Appendix Table A for the full list)

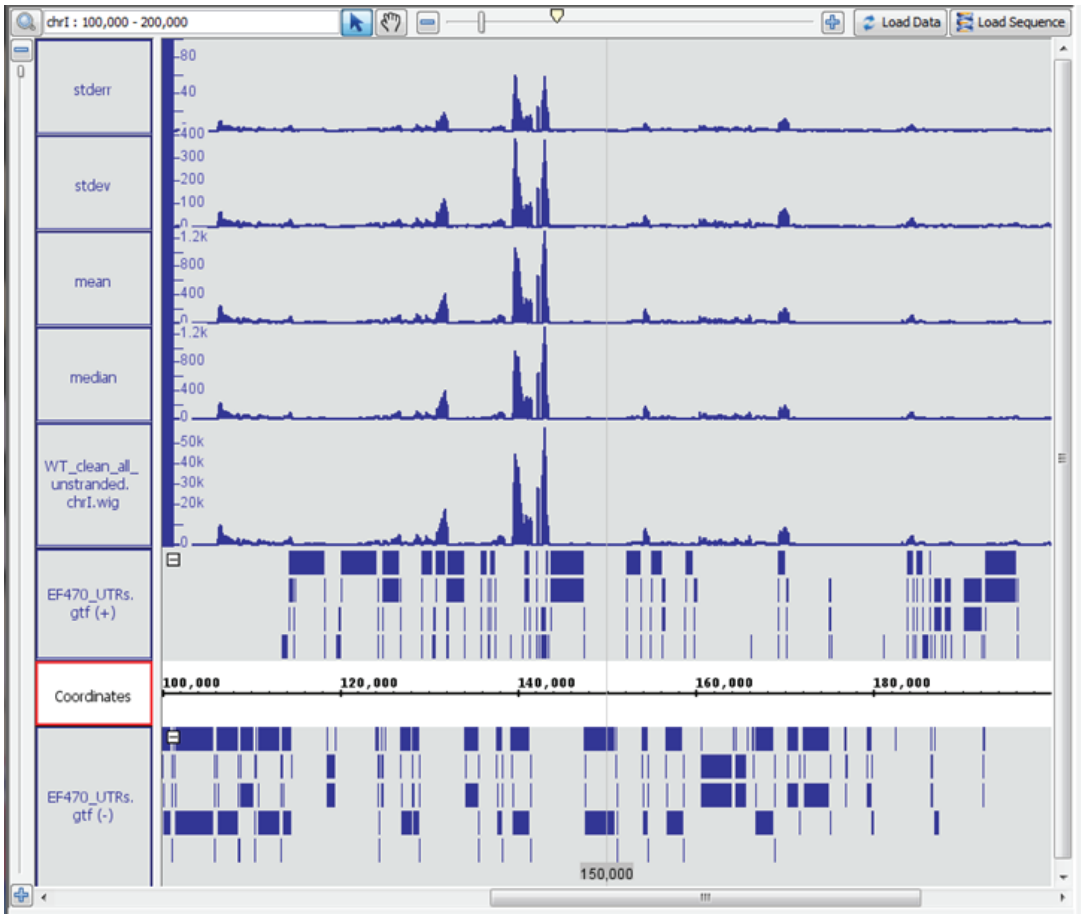
with flocculins, a group of proteins involved in the formation of clusters of yeast cells (Vidgren and Londesborough, 2011). However, the UAR did not have the necessary protein domains and repeats to be a functional flocculin. The third UAR studied in this chapter, chrV: 288,525–290,125, had very high homology to cell division control protein 4 (Cdc4); however, because of a stop codon within the frame of translation, any translated product would most likely be non-functional. Various methods were invoked throughout the examination of all three target UARs, including sorting by RNA-seq read depth and ORF length, BLAST searching (Altschul et al., 1990), and InterPro (Jones et al., 2014).

3.2 chrXII: 489,949–490,404

As reference, multiple versions of annotations are referred to throughout this section as this region was detected very early in the study, before further developments on gathering annotation. Table 3.1 lists the different sets of annotations and each of their contents.

Chromosome XII at 489,949–490,404 was the first interesting UAR found by exploring statistical figures of read depths. Means, medians, standard deviations, and standard errors of read depths per base were calculated over the 42 clean WT

Figure 3.1: A genomic region is shown in IGB against the EF470_UTRs set of annotations. Shown above the annoations are the read depths from the RNA-seq alignment (WT_clean_all_unstranded.chrI.wig) and then the values for all four statistics on read counts (from bottom to top: median, mean, stdev (standard deviation), and stderr (standard error)).

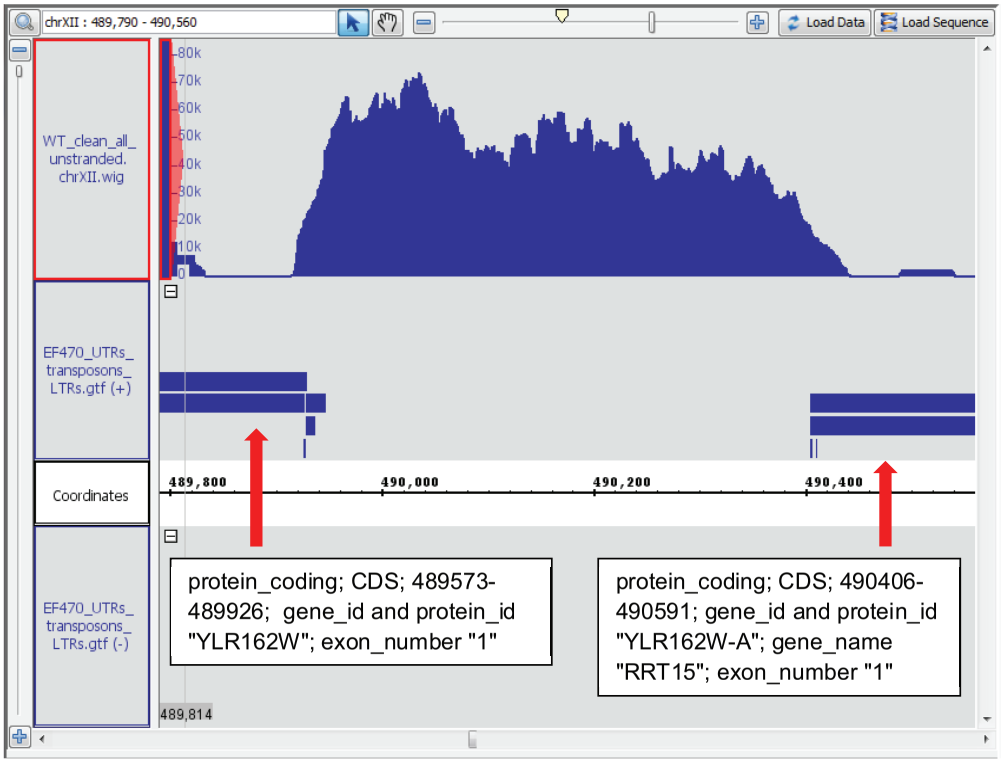


replicates. All four statistical values correlated with each other, as shown in Figure 3.1. Therefore, as the number of reads increased, so did the values for the four statistics. One region with highest number of reads stood out as a clear outlier: chrXII: 489,949–490,404.

Viewing the UAR in IGB showed protein-coding genes upstream and downstream of the region (Figure 3.2). As reference, the protein-coding region upstream of the UAR on the forward strand is YLR162W, annotated as a "putative protein of unknown function; overexpression confers resistance to the antimicrobial peptide

MiAMP1 and causes growth arrest, apoptosis, and increased sensitivity to cobalt chloride" (Kumar et al., 2011; Stephens et al., 2005) in SGD. YLR162W-A is downstream of the UAR, described as a "putative protein of unknown function identified by fungal homology comparisons and RT-PCR; identified in a screen for mutants with decreased levels of rDNA transcription" (Hontz et al., 2009; Kessler et al., 2003).

Figure 3.2: Region chrXII: 489,790–490,560 shown in IGB with UAR chrXII: 489,949–490,404 in the centre against the EF470_UTRs_transposons_LTRs set of annotations. The RNA-seq read alignment was performed with TopHat2.



To characterise this striking outlier UAR, multiple BLAST searches were performed. A BLASTN query against *S. cerevisiae* (taxid: 4932) resulted in three 100% matching regions (the third is the UAR itself) on chrXII. Table 3.2 lists all annotated features within all regions matched from BLASTN results. All features within these regions are rRNAs, strongly suggesting that the UAR may contain or be a rRNA itself, though experimental validation through differential and sucrose centrifugation, followed by ribosomal sequencing, would be necessary for certainty (Rivera et al., 2015). The top hits from BLASTN were yielded also by a TBLASTX search; however, "no significant similarity was found" by BLASTX.

With the final set of Primary Annotations, the UAR actually overlapped with a few annotations from data added later on in the study, eliminating it from the final pool of UARs (Figure 3.3). In a previous study that used high-resolution oligonucleotide tiling arrays, a meiotic unannotated transcript (MUT) was found at chrXII: 490,175–490,352, indicating that this region is active during the meiotic phase of the yeast reproductive cycle (Lardenois et al., 2011). However, the RNA-seq reads that mapped to the region previously with TopHat2 no longer appeared at this same region on chrXII, even with the most lenient, Near-Default, mapping. In terms of conservation, sequences of 6 other *Saccharomyces* species (*paradoxus*, *mikatae*, *kudriavzevii*, *bayanus*, *castelli*, and *kluyveri*) along with *cerevisiae* were aligned with the MULTIZ program (Blanchette et al., 2004). The multiple alignments were downloaded from the UCSC Genome Browser (Kent et al., 2002), which also provided calculated phastCons conservation scores for the multiple alignment (Siepel et al., 2005). For this partciular UAR, although the phastCons scores showed a maximum score of 1.0 across the entire previous UAR, there were actually no MULTIZ

Table 3.2: Three regions on chrXII were found by searching for the UAR sequence at chrXII: 489,949–490,404: itself and two others. The other two regions contained rRNA genes, described in this table.

Region on chrXII (genomic coordinates)	**Annotated feature within region**	**BLAST E-value**
451,981–452,436	RDN37-1 (location: 451,575–458,432; product type: rRNA; description: 35S rRNA, processed into the 25S, 18S, and 5.8S rRNAs)	0.0
451,981–452,436	RDN25-1 (location: 451,786–455,181; product type: rRNA; description: 25S rRNA, a component of the 60S subunit)	0.0
461,118–461,573	RDN25-2 (location: 460,923–464,318; product type: rRNA; description: 25S rRNA, a component of the 60S subunit)	0.0
489,949–490,404	N/A	

alignments amongst the 7 yeast species evident in this region.

After considering all final and updated Primary and Secondary Annotations, chrXII: 489,949–490,404 still has not been officially annotated as a rRNA, so there is scope for experimental validation. A study in 2004 detected that a small fraction of 25S-related rRNAs are polyadenylated in *S. cerevisiae*, making it plausible that this previous UAR may encode rRNAs that were detected by polyA-enriched RNA-seq (Kuai et al., 2004).

Figure 3.3: Region chrXII: 489,790–490,560 shown in IGB, a similar view to Figure 3.2. However, STAR produced the three read alignments, and with the latest set of Primary Annotations, chrXII: 489,949–490,404 does not exist as a single UAR. Instead, a meiotic unannotated transcript (MUT1050.1) located within the previous UAR was detected by a previous study (Lardenois et al., 2011). Starting with the demarcated Coordinates, Primary Annotations are given above (forward strand) and below (reverse strand). The track for polyadenylation sites for respective strands are shown next, followed by the three RNA-seq alignments. At the very top are Conservation Scores (0 to 1) for Multiple Alignments of 7 Yeast Genomes and the number of species aligned at each base below (0 to 7).

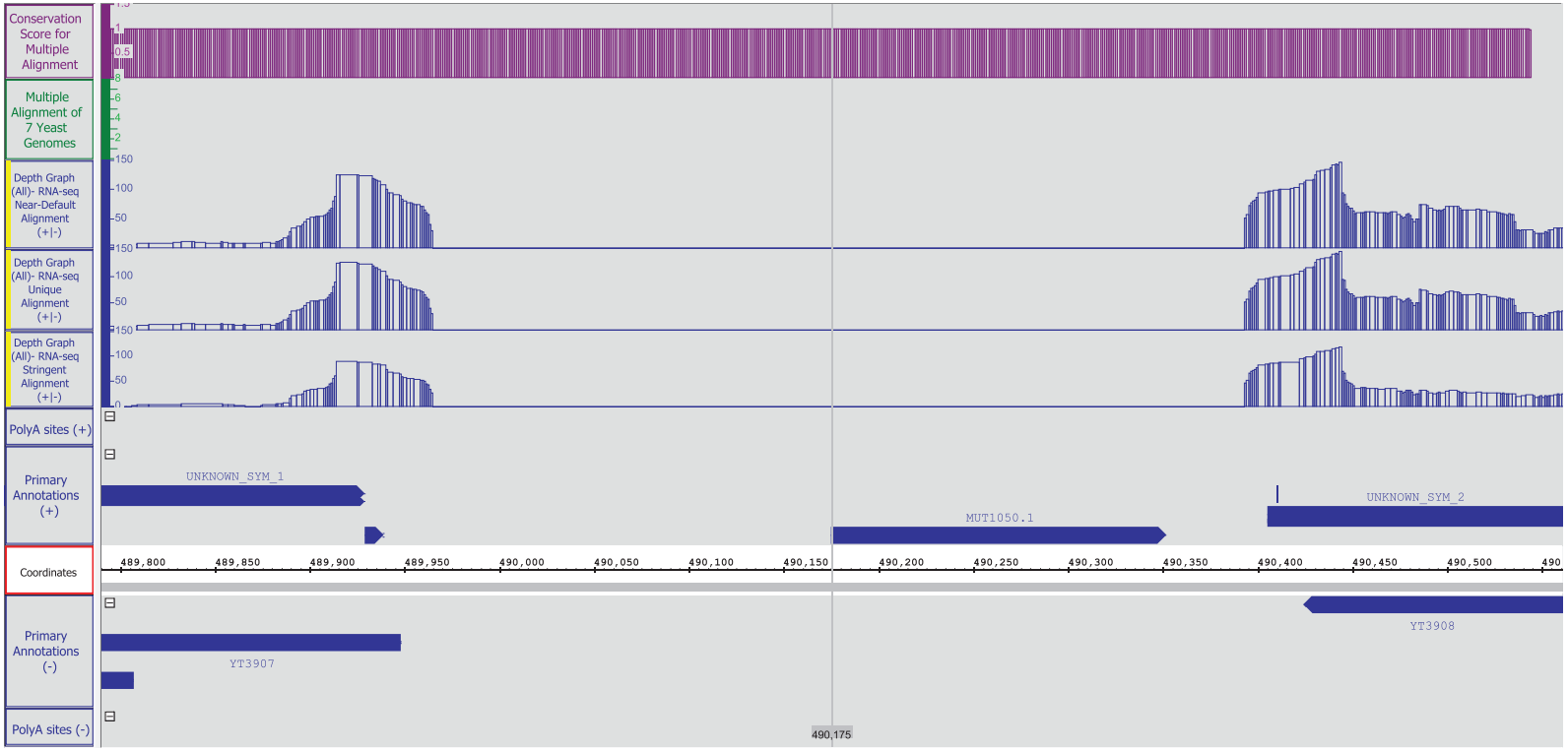


Table 3.3: The EF4.70 annotations with UTRs, transposons, and long-terminal repeats, a previous version of the Primary Annotations, and the TopHat 2 RNA-seq alignment yielded 4,534 UARs containin reads. These UARs were sorted by the total read depth, and the top 200 UARs were then subsequently sorted by the length of the longest ORF. The final top 10 UARs are listed by the length of the longest ORF.

Length of Longest ORF (bp)	**Un-Annotated Region**
426	chrXII: 218,908–220,666
389	chrIV: 804,518–806,444
249	chrI: 12,427–13,361
248	chrVIII: 542,263–543,006
248	chrI: 221,661–222,404
176	chrXIII: 282,856–284,036
167	chrVI: 255,428–258,853
166	chrIV: 1,164,764–1,166,960
166	chrI: 196,351–201,465
153	chrVIII: 1,898–2,669

3.3 chrI: 12,427–13,361

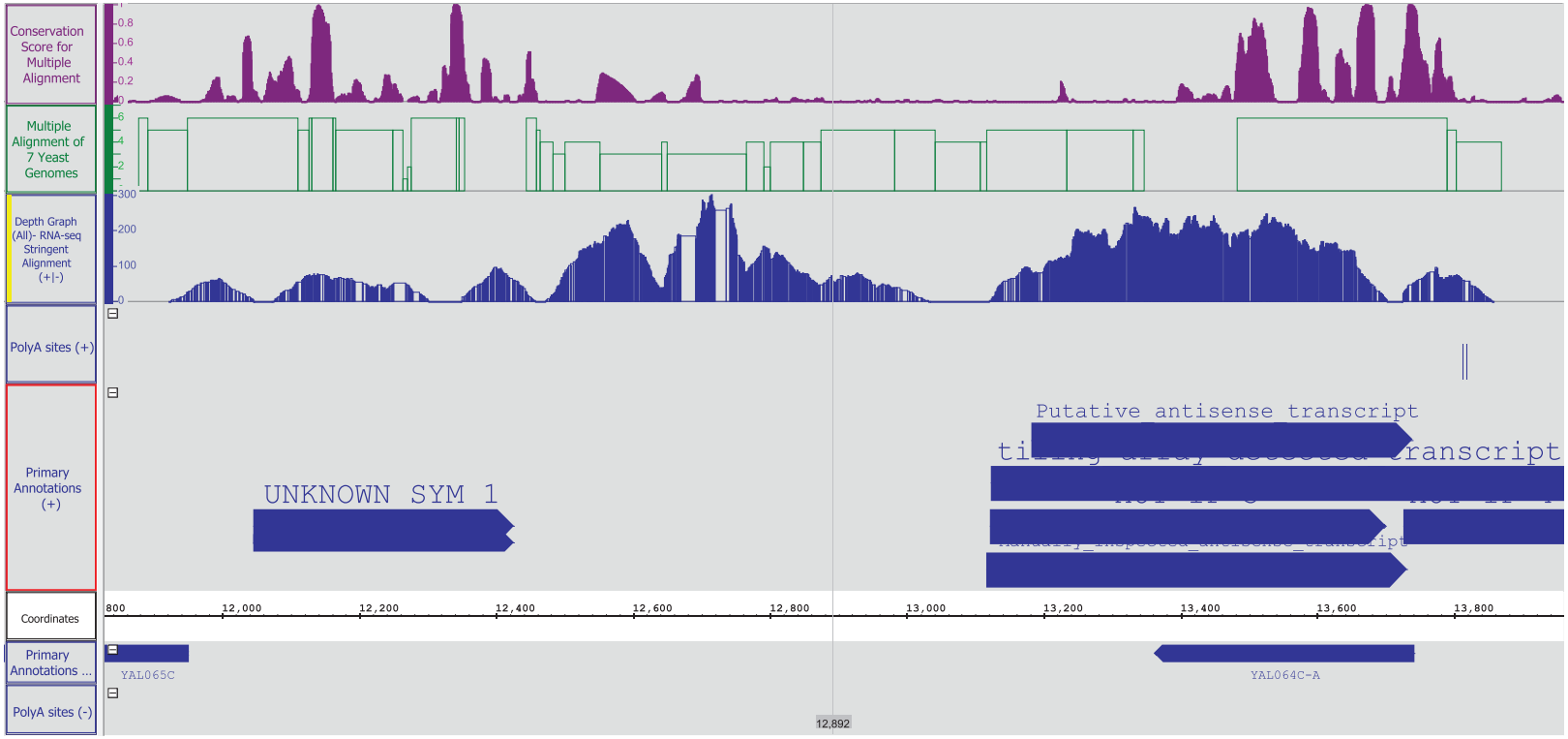
With the previous version of Primary Annotations, EF4.70 with UTRs, transposons, and long-terminal repeats, there were 4,534 un-annotated regions containing any mapped RNA-seq reads (55 UARs contained no reads). All of the 4,534 UARs were sorted by the total read depth (cumulative sum of read depth per base across the UAR, which differs from a total read count, the number of continuous 50-bp reads that overlap a particular region). The top 200 UARs of the sorted list were then sorted by the length of the longest ORF within the UAR (Table 3.3). Through BLASTX searches, the first two UARs unfortunately matched transposons and Gag-Pol fusion proteins, which are composed of Gag, the major virus coat protein, and Pol, an RNA-independent RNA polymerase (Ribas and Wickner, 1998). The third one, chrI: 12,427–13,361, matched flocculin proteins (Table C.1 in Appendix C).

Flocculins are proteins in brewer's yeast involved in "flocculation," a process

during which thousands of yeast cells aggregate (Vidgren and Londesborough, 2011). The aggregates are called "flocs," which settle rapidly to the bottom or rise to the surface in fermentation. Lectin-mediated adhesion is the most common process of flocculation (lectins are sugar-binding proteins). In this process, flocculin on the surface of a yeast cell binds to a mannose molecule in the cell wall of an adjacent cell. At least nine flocculin genes are known: FLO1, FLO5, FLO9, FLO10, FLO11, and four pseudogenes (Vidgren and Londesborough, 2011). The first four FLO proteins have a strong Flo1 phenotype with respect to the ability to bind sugar and inhibition by mannose. Flo11 is different since it is not directly involved with flocculation but, rather, flor formation in wines, filamentous growth, and solid surface adhesion. Moreover, Flo8 regulates the expression of the other FLO genes as a transcription factor (Vidgren and Londesborough, 2011).

Flocculins are unstable genes due to their long lengths (maximum of 4.6 kbp) and contain 10-20 tandem repeats of about 100 nucleotides (Vidgren and Londesborough, 2011). The tandem repeats are dynamic and change rapidly, and relocations of the repeats occur within and between flocculin genes. The repeats give flocculins the ability to participate in flocculation; the more repeats a flocculin possesses, the stronger its Flo1 phenotype. Genes with tandem repeats are difficult to map short reads to. Interestingly, all but FLO11 are located near telomeres. As a result, flocculin genes are susceptible to duplications, translocations, and deletions. In addition, telomeric silencing can repress transcription of these genes (Vidgren and Londesborough, 2011).

Figure 3.4: Region 11,827–13,958 on chrI displayed in IGB, showing the UAR chrI: 12,427–13,361 in the centre. Information tracks were organised in the same way as Figure 3.3. In this case, there was a distinct region of high RNA-seq read depth in the Stringent Alignment as well as a higher number of species in multiple alignments within the UAR. Conversely, the phastCons scores were not high at all across the entire UAR.



A closer examination of this UAR in IGB shows that it is surrounded by two nearby protein-coding regions, YAL065C and YAL064C-A (Figure 3.4). YAL065C is a putative protein of unknown function, has homology to FLO1, and is a possible pseudogene (Cherry et al., 2012). YAL064C-A is a putative protein of unknown function, its null mutant is sensitive to expression of the top1-T722A allele, and it is not an essential gene (Cherry et al., 2012). The top1-T722A allele stabilizes the Top1 cleavage complex, which consists of covalently attached Topoisomerase 1 and cleaved DNA (Andersen et al., 2015).

Within the UAR, the longest ORF was 249 amino acids. The longest ORF of all 6 reading frames that overlaps with this UAR was 535 amino acids long, extending from 11,569-13,174, which completely overlaps with the entire UAR region. An InterProScan search matched the region to only Flocculin type 3 repeats (Figure C.1 in Appendix C). The UAR was matched to Flo1, Flo5, and Flo9 by BLASTX (Table C.1), so the InterPro entry for each functional Flo protein was compared to that of the UAR. Flo1 does not have Flocculin type 3 repeats but possesses the PA14 (IPR011658) domain which forms an insert in yeast adhesins, amongst other proteins (Rigden et al., 2004), and the Flocculin repeats (IPR001389) (Figure C.2 in Appendix C). Interestingly, Flo5 and Flo9 did have the Flocculin type 3 repeats in addition to the PA14 domain and Flocculin repeats, showing that the UAR is most related to these two flocculin genes (Figures C.3 and C.4 in Appendix C).

The UAR possesses only the Flocculin type 3 repeats and not the critical PA14 domain or the Flocculin repeats that the functional FLO1, FLO5, and FLO9 have. Therefore, if there were a transcript derived from this UAR and it were then translated into a peptide, the peptide would most likely not be functional. It is plausible

that this UAR is a pseudogene, given the homology to other functional flocculin genes.

3.4 chrV: 288,525–290,125

The top 200 identified UARs were further refined with BLASTX searches against *S. cerevisiae*. Specifically, we used the BLASTX search results to identify UARS with the following characteristics:

- many high-quality hits with high similarity percentages and relatively low e-values
- many hits to other related yeasts, which may hint at conservation
- many hits to other eukaryotes outwith yeasts, alluding to wider conservation
- many hits to "unknown," "hypothetical," or "putative" entries for scope to further characterise these potential proteins

This resulted in four candidate UARs:

- chrVIII: 542,263–543,006
- chrI: 221,661–222,404
- chrV: 288,525–290,125
- chrI: 175,331–176,824

The first couple of UARs were very similar in sequence and returned almost identical BLASTX results. Furthermore, BLASTN search against *S. cerevisiae* returned sequences from other parts of the genome.

The third promising candidate, chrV: 288,525–290,125, yielded high similarity to cell division control protein 4 (Cdc4) in BLASTX results (Figure C.5 in Appendix

C). The alignment of Cdc4 against the continuous region shows homology and indicates also where the stop codon is located against the functional protein (Figure C.7 in Appendix C). Cdc4 assists target cell cycle regulators in ubiquitin-mediated proteasomal degradation (Goh and Surana, 1999). All BLASTX hits aligned to the 3' end of the UAR, even past the region's longest ORF of 126 aa at 289,528–289,905. The UAR was extended at the 3' end to encompass a 296-aa ORF adjacent to the first ORF at 289,908–290,799. The extended UAR at 288,525–291,000, 200 bp past the second ORF, was searched via BLASTX (Figure C.6 in Appendix C). Most of the hits matched the continuous region consisting of both ORFs. Consequently, if the region spanning both ORFs were transcribed and translated, it would produce a truncated, perhaps non-functional, version of Cdc4. Therefore, this UAR is most likely part of a pseudogene, similar to the previous preliminary UAR target with homology to flocculins.

The three preliminary target UARs described here were flagged as interesting regions primarily through RNA-seq and manual searching with BLAST and Interpro. However, there was a need for more automated and comprehensive approaches. In the following stages of the study, techniques such as snoRNA prediction and proteomics were integrated with the RNA-seq methods to investigate whether there were possibilities of other ncRNAs, peptides, or proteins being expressed. The next section describes some of the aforementioned methods in further detail, while the next chapter describes several additional tools that were developed to facilitate analysis of the UARs.

3.4.1 RNA-sequencing

The following sections outline the main steps in an RNA-seq experiment according to the High Sample Protocol from the Illumina TruSeq RNA Sample Preparation v2 Guide (Illumina, 2014).

Purification and Fragmentation of mRNA

Total RNA is extracted from the organism and then diluted. RNA purification beads are mixed with the total RNA. The mixture is heated and incubated. The supernatant is then discarded. The remainder is washed with bead washing buffer and resuspended in elution buffer. Messenger RNA is eluted from the beads, and bead binding buffer is added to the mixture, which is incubated for rebinding. The mixture is washed with bead washing buffer and the Elute, Prime, Fragment mix, which contains random hexamers for priming, is added. The mixture is then aliquoted to wells on an RNA fragmentation plate, which is heated to elute, fragment, and prime the RNA.

First Strand cDNA Synthesis

The RNA fragmentation plate is placed on a magnetic stand to ensure the beads are bound to the sides of the well, and from each well, the supernatant (containing fragmented and primed mRNA) is transferred to a well on a Hardshell plate. The First Strand Master Mix and reverse transcriptase are mixed in to each well on the Hardshell plate, which is now considered the cDNA plate. The cDNA plate is then placed and run in a pre-programmed thermal cycler to create the first strand cDNA.

Second Strand cDNA Synthesis

To each well of the cDNA plate, Second Strand Master Mix is added and mixed in. AMPure XP Beads are added to each well of a new MIDI plate, now considered the cDNA clean up plate (CCP). All content from each well of the cDNA plate are transferred to a corresponding well of the CCP. The CCP is placed on the magnetic stand, and the supernatant from each well is discarded. Ethanol is added to each well, and the supernatant is discarded again (repeat once). The CCP is left to dry, and Resuspension Buffer is mixed into each well. The cDNA plate is placed on a magnetic plate, and the supernatant (ds cDNA) is transferred to a new MIDI plate, that is now considered the insert modification plate (IMP).

End Repair

To each well in the IMP containing ds cDNA, End Repair Mix is added and mixed in. The plate is then shaken, centrifuged, heated, and incubated. The IMP plate is placed on a magnetic stand, and the supernatant of each well is discarded. Each well is washed twice with ethanol, and then the plate is left to dry. The dry pellet in each well is mixed with Resuspension Buffer. The plate is placed again on the magnetic stand, and the supernatant from each well is transferred to a corresponding well of a new MIDI plate, now considered the adapter ligation plate (ALP).

Adenylation of 3' Ends

To each well of the ALP, A-Tailing Mix is mixed in. The ALP is heated and incubated in two cycles.

Ligation of Adapters

Ligation Mix is mixed in to each well of the ALP. The ALP is heated and incubated. To inactivate the Ligation Mix, the Stop Ligation Buffer is added to each well. AMPure XP Beads are added to each well, and the plate is placed on a magnetic stand. Supernatant from each well is discarded, each well is washed twice with ethanol, and the plate is left to dry. To each well, Resuspension Buffer is mixed in, and the plate is placed on the magnetic stand again. Supernatant from each well is transferred to a corresponding well of a new MIDI plate, now considered the clean up ALP plate (CAP).

To each well of the CAP, AMPure XP Beads are mixed in. The CAP is placed on a magnetic stand, and the supernatant of each well is discarded. Each well is washed twice with ethanol, and the plate is left to dry. To each well, Resuspension Buffer is mixed in. The CAP is placed on a magnetic stand, and the supernatant from each well is transferred to a corresponding well of a new Hardshell plate, now considered the PCR plate.

DNA Fragment Enrichment

To each well of the PCR plate, PCR Primer Cocktail and PCR Master Mix are added. The PCR plate is run on a pre-programmed thermal cycler for 15 cycles. AMPure XP Beads are added to each well of a new MIDI plate, now considered the clean up PCR plate (CPP). Contents of each well from the PCR plate are transferred to the CPP and mixed. The CPP is placed on a magnetic stand, and the supernatant of each well is discarded. Each well is washed twice with ethanol and left to dry. Dried pellets are mixed with Resuspension Buffer, and the plate is

placed on a magnetic stand. The supernatant of each well is transferred to a new Hardshell plate, now considered the Target Sample Plate 1 (TSP1).

Library Validation

Quantitation of DNA library templates is determined by qPCR. Libraries are prepared from nucleic acid sequences that are amplified to yield clonal clusters, which are sequenced in parallel. The quality of data and total data output depend on the density of clonal clusters. Therefore, optimum cluster densities across every lane of the flow cell should be determined using quantitation from qPCR.

Quality control regarding the size and purity of the sample may be performed, for example, by applying the sample onto a DNA-specific chip (Agilent DNA 1000). A band at about 260 bp should appear for single-read libraries.

Normalisation and Pooling of Libraries

Sample library from each well of the TSP1 is transferred to a corresponding well of a new MIDI plate, now considered a diluted cluster template. The concentration of sample library in each well of the DCT is normalised using a solution of Tris-HCl 10 mM, pH 8.5 with 0.1% Tween 20. For non-pooled libraries, this is the final step before cluster generation.

For pooled libraries, before cluster generation an aliquot of each normalized sample library is transferred from the DCT plate to a single well of a new Hardshell plate, now considered the pooled DCT plate.

Cluster Amplification

The DNA library is loaded onto a flow cell, which is coated with 2 oligonucleotides called 'p5' and 'p7' (Illumina, 2016). As DNA fragments flow across the oligo lawn, their adapter ends hybridise with and bind to complimentary oligos. The opposite end of a DNA fragment that has ligated bends over and connects to another complementary oligo on the flow cell, forming a bridge. Clustering consists of repeated cycles of denaturation and extension yields local amplification of single DNA fragments into clonal clusters across the flow cell.

Sequencing

Fluorescent nucleotides are added to the flow cell for the first base to bind (Illumina, 2016). An image is taken of the flow cell, and each cluster's emission is recorded. The four different bases fluoresce at different wavelengths, the method of identification. These steps are repeated 'n' times for a read that has an 'n' length of bases.

Modifications for Stranded Sequencing

The above protocol produces unstranded RNA-seq reads. For strand information, the Illumina TruSeq Stranded mRNA Sample Preparation Guide specifies the use of the Second Strand Marking Master Mix (Illumina, 2013). The Mix contains dUTP instead of dTTP, which marks the second strand for degradation with uracil-DNA glycosylase (Zhang et al., 2012).

Modifications for Paired-End Sequencing

The aforementioned protocol can be modified such that adapters for 5' and 3' ends of each DNA fragment are distinct sequences (Illumina, 2011). Reads are generated from both the 5' and 3' ends, and the distance between each paired read is known. Therefore, reads can be aligned to repetitive regions more accurately (Illumina, 2016).

Chapter 4

Proteomics

4.1 Introduction

This chapter describes how un-annotated regions in *Saccharomyces cerevisiae* were characterised by integrating two orthogonal high-throughput methods: RNA-seq and proteomics. The RNA-seq data provide a snapshot of the expression of the yeast genome, while a high-quality SILAC-based proteomics experiment gives a deep sampling of peptides and proteins present in the organism. The objective was to find open reading frames within un-annotated regions containing RNA-seq read alignments with corresponding peptides that were detected in the proteomics data. The raw proteomics data were searched against the hypothetical peptides coded by these open reading frames, resulting in the detection of two UAR ORFs that were found to be expressed at both the RNA transcript and peptide levels. The effects of the size of the proteomics sequence database on the peptide identification results were also explored.

4.2 Heat Stress Proteomics Dataset

4.2.1 Experimental Protocols for Data Production

The following sections describe the experiment performed by collaborators in the Pedrioli Laboratory at the University of Dundee.

Culture and SILAC Labelling

Strains of BY4741 were grown in yeast extract peptone dextrose to mid-log phase for at least four doublings (Tyagi and Pedrioli, 2015). Cultures were exposed to 37 degrees C for at least four doublings under the heat-stress treatment. Strains to be SILAC labelled were grown in synthetic complete media without lysine or arginine, supplemented by proline, 13C6,15N4-arginine, and 13C6,15N2-lysine. Unlabelled cultures were grown in synthetic complete media, supplemented by proline, unlabelled arginine, and unlabelled lysine. SILAC labelled (heavy) and unlabelled (light) cultures were grown for at least four doublings. Equal OD600 of labelled and unlabelled cultures were mixed for further processing after harvesting (Tyagi and Pedrioli, 2015).

Extraction, Digestion, and Fractionation

Harvested yeast cells were treated with 1.85 M NaOH and 7.6% (v/v) -mercaptoethanol for 10 min on ice and 50% (w/v) trichloroacetic acid at equal volume for 20 min on ice (Tyagi and Pedrioli, 2015). Proteins were precipitated, pelleted, and washed with acetone. The FASP method was used for tryptic digestion. C18 MacroSpin

columns were used to purify the digested peptides (Tyagi and Pedrioli, 2015). Isoelectric focusing was used to purify peptides into 12, 17, or 13 fractions, yielding three sets or replicates with an OffGel Fractionator (Tyagi and Pedrioli, 2015).

LC-MS/MS

Peptides were re-suspended in CF3COOH (Tyagi and Pedrioli, 2015). Online reverse phase liquid chromatography was used on the Ultimate3000 uHPLC system. Coated-tip fused silica columns were packed with C18 silica beads for a length of 50 cm. Peptides were resolved with a 250 nL/min gradient of buffer B (0.1% (v/v) HCOOH, 90% (v/v) CH3CN, 3% (v/v) DMSO) in buffer A (0.1% (v/v) HCOOH, 2% (v/v) CH3CN, 3% (v/v) DMSO) ranging from 2% to 35% over 240 min for fractionated samples at 45 degrees C. An LTQ-Orbitrap Velos Pro mass spectrometer was directly coupled to the chromatography apparatus. The method of top-15 data dependent acquisition by collision-induced fragmentation was used. At the FT-MS resolution of 60,000, MS1 scans were taken in profile mode. IonTrap Rapid Scan Rate took MS/MS scans in centroid mode. The following configurations were used:

- precursor ion intensity threshold for triggering fragmentation = 500 arbitrary units
- dynamic exclusion = enabled
- repeat count = 1
- repeat duration = 30 s
- exclusion list size = 500
- exclusion duration = 90 s

Table 4.1: The proteomics experiment was performed three times under the following conditions (Tyagi and Pedrioli, 2015).

Replicate Name	**Condition**	**Number of Fractions**
rep1	30 degrees C heavy, 37 degrees C light	12
rep2	30 degrees C light, 37 degrees C heavy	17
rep3	30 degrees C heavy, 37 degrees C light	13

- preview mode for FT-MS master scans = enabled
- monoisotopic precursor selection = enabled
- charge state screening with +1 rejection = enabled

4.2.2 Evaluation of Proteomics Data

Tyagi and Pedrioli (2015) states that there were 4,663 proteins detected from the proteomics dataset by ProteinProphet (Nesvizhskii et al., 2003) at 1% FDR. Of these, 4,612 proteins were identified by at least one peptide above the 1% FDR threshold from PeptideProphet (Keller et al., 2002), covering 68% of the predicted *S. cerevisiae* proteome.

The quality and performance of the proteomics data were assessed in Tyagi and Pedrioli (2015) also by comparing the number of proteins detected against all sequences present in:

- the *Saccharomyces* Genome Database, the central repository of all annotations available for yeast
- another proteomics experiment in which all annotated ORFs were attached to high-affinity epitope tags for immunodetection (Ghaemmaghami et al., 2003)
- the PeptideAtlas repository, which hosts a collection of peptides detected in

Table 4.2: The number of proteins detected in the proteomics dataset against the number of proteins in *Saccharomyces* Genome Database, a tag-based proteomics method, and PeptideAtlas to illustrate performance. Reproduced from Tyagi and Pedrioli (2015) with permissions.

Category	**SGD**	**Heat Stress Proteomics Dataset (% of SGD)**	**Tag-based proteomics from Ghaemmaghami et al. (2003)**	**PeptideAtlas (King et al., 2006)**
Verified	4,939	4,198 (85)	4,048	4,197
Uncharacterised	853	401 (47.01)	409	402
Dubious	810	4 (0.49)	41	11
Transposable element	89	4 (4.49)	3	55
Pseudogene	26	5 (19.23)	3	3
Total	6,717	4,612 (68.66)	4,504	4,668

tandem mass spectrometry experiments across multiple species (King et al., 2006)

The numbers of proteins detected in the aforementioned experiments and repositories are listed in Table 4.2. There were 108 more proteins detected in the heat stress experiment than in the tag-based approach. In addition, the total number of proteins identified in this single heat stress experiment was only 56 (1.2%) proteins smaller than the entire collection of proteins detected over all yeast experiments contained in PeptideAtlas. This difference demonstrates that as an individual experiment, the heat stress proteomics study provided extensive coverage over the yeast proteome, rendering it a comprehensive dataset to search the un-annotated regions against.

4.3 Sequence Database Construction

Proteomics data are searched against a FASTA-formatted database of peptide and/or protein sequences of interest. The database is entirely customisable and can include hypothetical amino acid sequences, which was advantageous in the analysis of un-annotated regions. Figure 4.1 shows the process by which the database was built. Firstly, the full-length DNA sequences of all chromosomes in yeast were acquired from the *Saccharomyces* Genome Database. The sequences were 6-frame translated with the Transeq (Rice et al., 2000) program to identify all possible open reading frames over the entire genome. The translated chromosomes were contained in separate .fasta files, one per chromosome, but were concatenated into a single file for convenience. The translated chromosomes, along with coordinates of all 2,636 un-annotated regions and DNA coordinates of all ORFs, were inputs for the *extract_orfs_for_fasta_list.py* script. The script processes all inputs by chromosome. For each UAR, the .tab file for all ORFs is searched for the ORFs that completely overlap with the specific UAR. Then, after converting the DNA coordinates into peptide coordinates for each ORF, the peptide sequence for that particular ORF is extracted from the single .fasta file of translated chromosomes. After all of the sequences were collected, an entry per UAR ORF was written to a .fasta file that included the chromosome, start, end, translation frame, and peptide sequence. Since the smallest translated ORF reported in literature across all species is 6 amino acids (Andrews and Rothnagel, 2014), the entire pool of hypothetical peptides were filtered for those at least 6-aa in length, contributing 22,852 sequences to the database.

One way to assess the coverage and quality of a proteomics dataset is to search

Figure 4.1: Schematic diagram of how the sequence database for the proteomics analysis was constructed. Software programs are in bold and scripts are in italics.

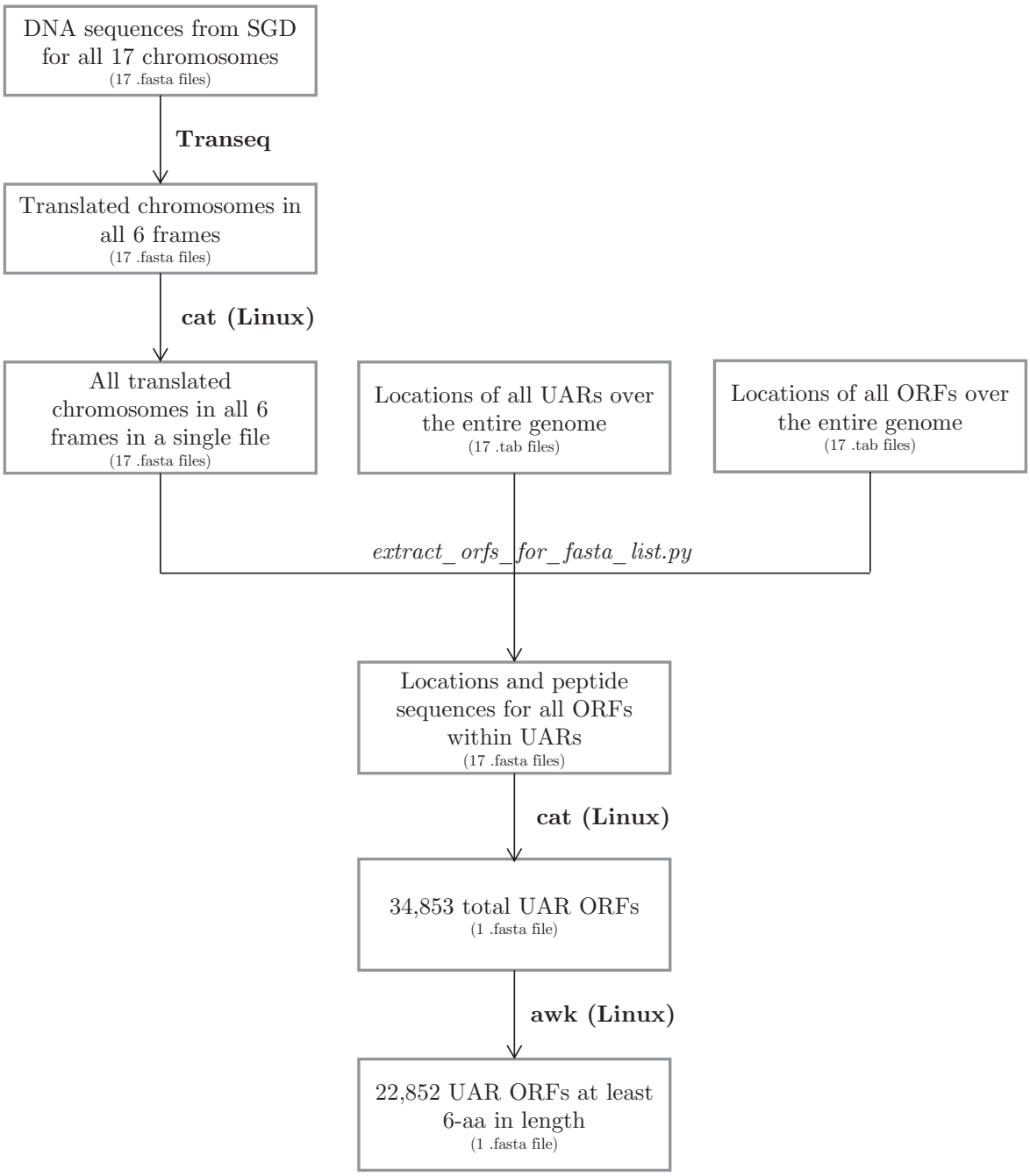


Table 4.3: Contents of the FASTA-formatted peptide/protein sequence database.

Category	**Number of Sequences**
SGD known genes	6,603
Reversed SGD known genes	6,603
UAR ORFs at least 6-aa long	22,852
Reversed UAR ORFs at least 6-aa long	22,852

the mass spectra against known peptide and protein sequences to ensure that these sequences were detected with high probabilities. Therefore, the 6,603 known SGD genes were added to the database. False discovery rate (FDR) are determined thresholds used to derive a set of Peptide-Spectrum Matches (PSMs) in proteomics (Choi and Nesvizhskii, 2008). The estimation of FDRs relies on measuring the detection of decoys, or false peptide/protein sequences. False peptide/protein sequences can be created by reversing the 6,603 known SGD genes and inserting these into the sequence database. As a summary, Table 4.3 lists the entire contents of the sequence database. Reversed sequences of SGD known genes and UAR ORFs were included in the database mainly to allow the software to estimate the false positives in a set of peptide spectrum matches at a given score threshold. Therefore, any decoy identifications are discarded.

4.3.1 Unit Testing

Python unit tests in the *extract_orfs_for_fasta_list.py* script ensured that all ORFs within an UAR, and only the ORFs that entirely overlapped, would be detected. Unit tests were written to verify that the conversion of DNA sequence coordinates to peptide sequence coordinates, which was crucial for determining the exact corresponding peptide for an ORF.

4.4 Proteomics Analysis Procedure

The proteomics analysis work flow per replicate is shown in Figure 4.2. The third-party software programs and scripts invoked are described in Tables 4.5 and 4.4, respectively. For reference, the file structure adopted for the proteomics analysis is illustrated in Figure 4.3 to aid in the understanding of inputs and outputs in the work flow. Automating the proteomics analysis process would facilitate the use of this pipeline.

Figure 4.2: Work-flow diagram of the proteomics data analysis. Scripts are italicised, and software programs are embolden, details of which in Table 4.4, and Table 4.5, respectively.

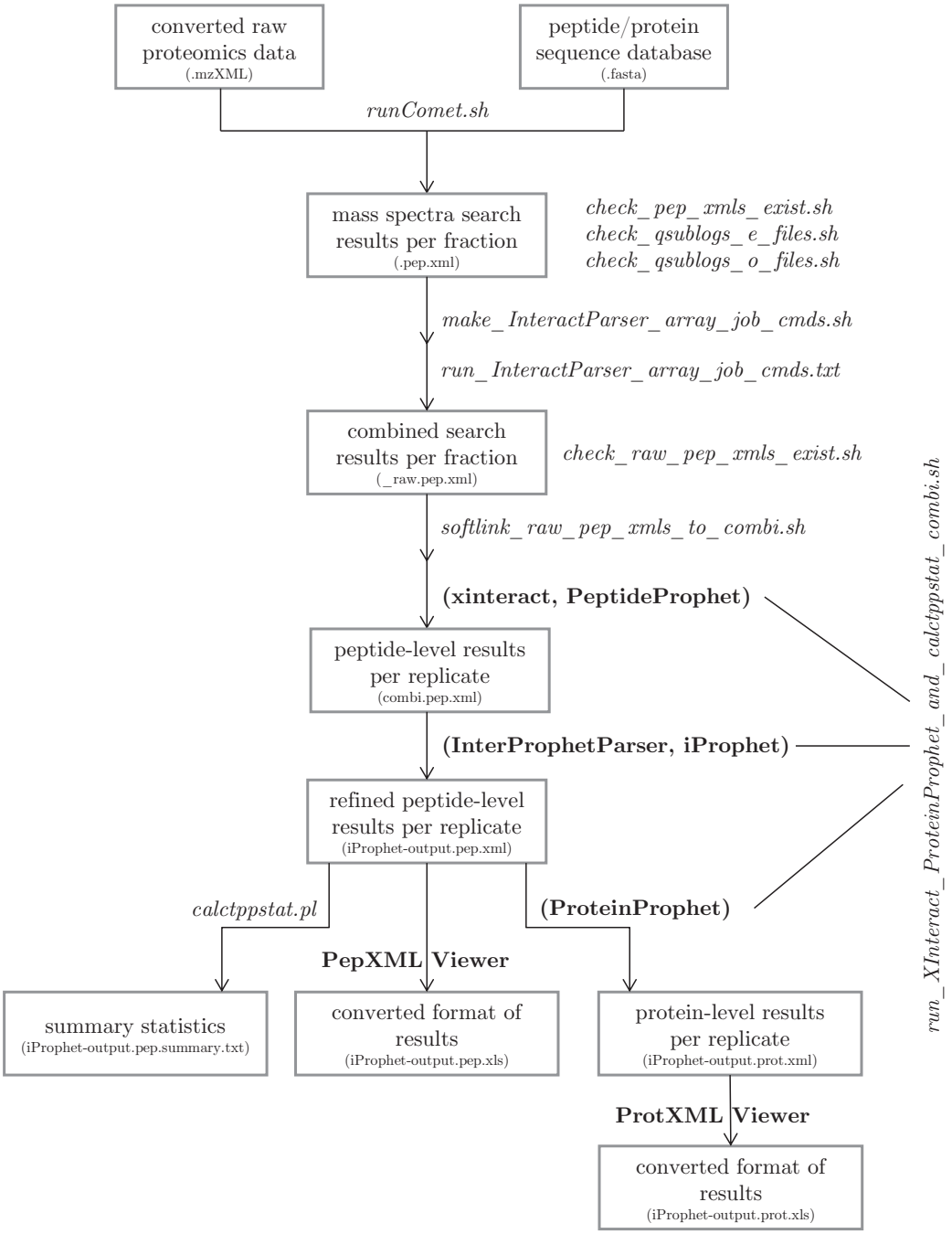


Table 4.4: Functions of scripts run in the proteomics analysis pipeline shown in 4.2. Table 4.5 contains further details on programs called by these scripts.

Script/File	**Source**	**Description/Function**
runComet.sh	TPP, Comet	divides each .mzXML file into multiple files and performs Comet searches of tandem mass spectra against a sequence database
check_pep_xmls_exist.sh	author	lists the name of each fraction and its pep.xml files
check_qsublogs_e_files.sh/ check_qsublogs_o_files.sh	author	lists all the qsublog error ("e") or output ("o") files and their file sizes for detection of any aberrations, which can indicate an incomplete run
make_interactParser_array_job_cmds.sh	author	Arguments required: the Comet directory, the to_search.txt path, and the output file name (e.g. run_InteractParser_array_job_cmds.txt). Produces a .txt file of commands to run InteractParser.
run_InteractParser_array_job_cmds.txt	author	a list of commands for a Sun Grid Engine qsub array job that runs the InteractParser program
check_raw_pep_xmls_exist.sh	author	lists the name of each fraction and its _raw.pep.xml file
softlink_raw_pep_xmls_to_combi.sh	author	softlinks _raw.pep.xml files to the combi sub-directory
run_XInteract_ProteinProphet_and_calctppstat_combi.sh	author, TPP	runs xinteract, InterProphetParser, ProteinProphet, and calctppstat.pl

Table 4.5: Functions of third-party software programs for proteomics analysis and the relevant parameters used with each.

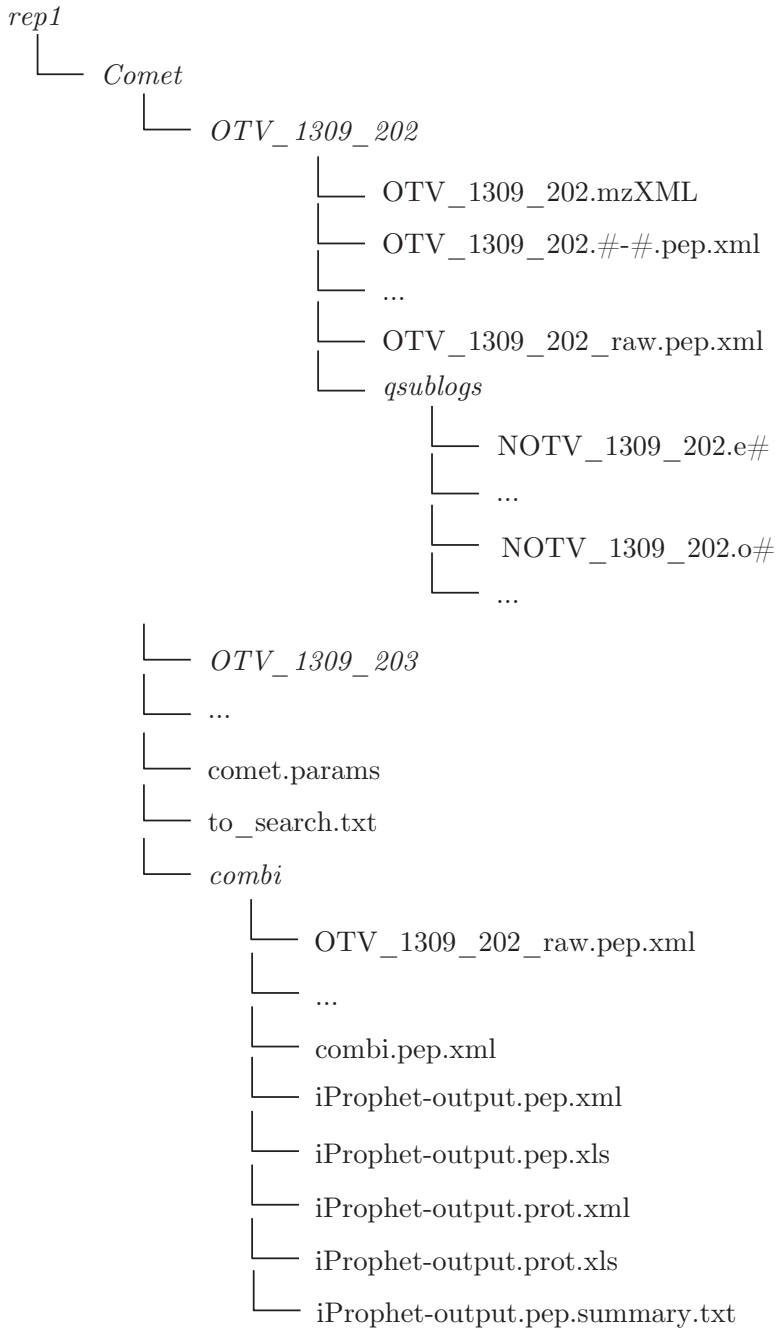
Program	**Function**	**Parameters/Options**
Comet	searches tandem mass spectra of peptides against a sequence database	in comet.params file: database_name = "RandomisedDB_6_aa_min.fa" (Table 4.3), variable_mod1 = 8.01419892 K 0 3 (mass difference from isotope labels, lysine, 0 = all permutations of modified and unmodified residues are analysed, 3 = maximum of 3 modifications per peptide), variable_mod2 = 10.008252778 R 0 3 (see variable_mod1 but for arginine)
InteractParser	combines all .pep.xml files into a single _raw.pep.xml file	_raw.pep.xml *-*.pep.xml
xinteract	command-line wrapper for several TPP programs	see PeptideProphet, iProphet, and ProteinProphet
PeptideProphet	performs peptide-level searches and provides probabilities for identifications for all files ending in _raw.pep.xml to output in a single combi.pep.xml file	-OAPd -drev_ -PPM -Ncombi.pep.xml *_raw.pep.xml (O = indicates to XInteract that these are options for PeptideProphet, A = accurate mass binning, P = non-parametric model, d = report decoy hits with a computed probability based on the model learned, -drev_ = use "rev_" decoy hits to build the negative distribution, -PPM = use parts per million instead of Daltons in the accurate mass model)
InterProphetParser	a wrapper script that runs iProphet on the PeptideProphet results in combi.pep.xml and outputs to the iProphet-output.pep.xml	combi.pep.xml iProphet-output.pep.xml
iProphet	performs statistical refinement of PeptideProphet probabilities	(see InterProphetParser)

ProteinProphet	identifies proteins from PeptideProphet or iProphet results in the .pep.xml file and produces probabilities for identifications in the iProphet-output.prot.xml file	iProphet-output.pep.xml iProphet-output.prot.xml IPROPHET (input is from iProphet) GROUPWTS (check peptide's total weight, instead of the actual weight, in the Protein Group against the threshold) NORMPROTLEN (normalise the number of sibling peptides against the threshold)
calctppstat.pl	creates a summary of the results and statistics for	-i iProphet-output.pep.xml (input file) -d rev_ (decoy string)
PepXML Viewer	web interface for viewing pep.xml files and exporting to .xls format ('TPP v0.0 TRUNK (DEV) rev 0, Build 201405152128 (linux))	N/A
ProtXML Viewer	web interface for viewing prot.xml files and exporting to .xls format ('TPP v0.0 TRUNK (DEV) rev 0, Build 201405152128 (linux))	N/A

In the first instance, the raw proteomics data, converted to .mzXML format, were searched against the sequence database described in Section 4.3 with the Comet (Eng et al., 2013) search engine. Results from the mass spectra searches were produced in multiple .pep.xml files per fraction. To confirm that searches for all fractions were performed successfully, the *check_pep_xmls_exist.sh*, *check_qsublogs_e_files.sh*, and *check_qsublogs_o_files.sh* scripts were written for the user to detect any aberrations in the existence of files or the file sizes of error and output log files. For each fraction, all .pep.xml files were then combined into a single _raw.pep.xml file with the InteractParser (Deutsch et al., 2010) program, which was run by the commands in *run_InteractParser_array_job_cmds.txt*. Afterwards, all _raw.pep.xml files for each fraction were then symbolically linked in the combi directory.

The next script,*run_XInteract_ProteinProphet_and_calctppstat_combi.sh*, runs multiple third-party programs in succession. To begin with, the wrapper program xinteract (Deutsch et al., 2010) runs PeptideProphet (Keller et al., 2002) to compute the probabilities of peptide assignments to mass spectra from the Comet searches in order to delineate between correct and incorrect assignments. These results were populated in a single combi.pep.xml file. Next, InterProphetParser (Deutsch et al., 2010) was executed to run iProphet (Shteynberg et al., 2011), which performed a statistical refinement of PeptideProphet probabilities, written in the iProphet-output.pep.xml file. ProteinProphet (Nesvizhskii et al., 2003), which computes probabilities for protein identifications, was run on the iProphet results, and the output was then stored in the iProphet-output.prot.xml file. As a final step, the script calctppstat.pl (Deutsch et al., 2010) was run to report a statistical summary of the peptide and

Figure 4.3: Structure of directory where the proteomics analysis files are stored. Directories (folders) are italicised, hash symbols represent numbering, and ellipses indicate the presence of more items with similar content as the item directly above.



protein identifications, including 1% false-discovery rate thresholds for the probabilities. The summary was written in the iProphet-output.pep.summary.txt.

The peptide-level and protein-level identifications were displayed in the PepXML and ProtXML Viewers (Deutsch et al., 2010), respectively. The Viewers are also capable of exporting the .xml result files into .xls (comma-separated) for a more accessible format to perform downstream analysis.

4.4.1 Proteomics Data Processing Parameters

The Trans-Proteomic Pipeline was used to process data (Pedrioli, 2010), with which raw data files were converted to mzXML format (Pedrioli et al., 2004). Comet was used to search mzXML files were searched against the databases described later in the text. Configurations for the search were as follows:

- Static modification:
  - carboxyamidomethylation of Cys (57.022 Da)
- Variable modification:
  - 13C6, 15N2-Lys (8.01419892 Da)
  - 13C6, 15N4-Arg (10.008252778 Da)
  - oxidation of Met (15.99491463 Da)
- Maximum of missed cleavages by tryptic digestion = 2
- MS error mass tolerance= 25 ppm
- MS/MS error mass tolerance = 0.4

PeptideProphet was used to analyse and evaluate peptide identification probabilities (Nesvizhskii et al., 2003), while ProteinProphet was used for protein identifications (Keller et al., 2002). iProphet was then applied to improve peptide identification rates and error estimation (Shteynberg et al., 2011). One consideration of the configurations listed is that the settings would not allow for peptides in which there were 3 or more missed cleavages. It would be beneficial to to able to measure the efficiency of the trypsin digestion to determine whether that step should be improved and how much sample was excluded; however 85% of verified SGD proteins were reported by Tyagi and Pedrioli (2015), which is higher than the other two studies compared against (Ghaemmaghami et al., 2003; King et al., 2006). Therefore, the maximum of 2 missed cleavages yields results of commensurate quality to other studies with high rates of protein detection.

4.5 Proteomics Search Methods

Five total search methods were invoked in this project. Three searches were performed on databases for which un-annotated regions open-reading frames were filtered by a minimum length. As expected, increasing the minimum length decreased the total number of ORFs in each database the proteomics data were searched against. The objective for the other two searches were to curate a database containing a set of known SGD protein-coding genes to serve as a set of likely candidates to be detected by proteomics searching, resembling a group of true positives.

4.5.1 Characterisation of Un-Annotated Region Open Reading Frames

Boxplots of the lengths of all un-annotated region open reading frames per chromosome are shown in Figure 4.4. Across all 16 non-mitochondrial and 1 mitochondrial chromosomes, the median lengths ORFs are under 50 bp in length, with the mitochondrial chromosome having the shortest median. It should be noted that these lengths are indicated in bp, but since ORFs are translated to their theoretical peptide sequences, the numbers given can be divided by 3 to depict the length of hypothetical peptides or proteins. Although the UAR ORFs are typically short, there is a significant tail that extends out to lengths that approach the lengths of protein-coding ORF in yeast (median prot ORF =1071bp). Therefore, it is feasible that a new peptide or protein may be detected by proteomics within this set of translated ORFs.

Figure 4.4: Boxplots of distributions of the lengths of un-annotated region open reading frames across all chromosomes.

The distribution of UAR ORFs amongst all chromosomes is shown in Figure 4.5. It is striking that the mitochondrial chromosome has the highest density of UAR ORFs compared to all of the other chromosomes; however, this may be explained by the low number of genes on the chromosome. In the proteomics search databases, there are only 28 protein-coding genes on the mitochondrial chromosome even though it is 85,779 bp long. Therefore, there are long un-annotated regions which produce many ORFs.

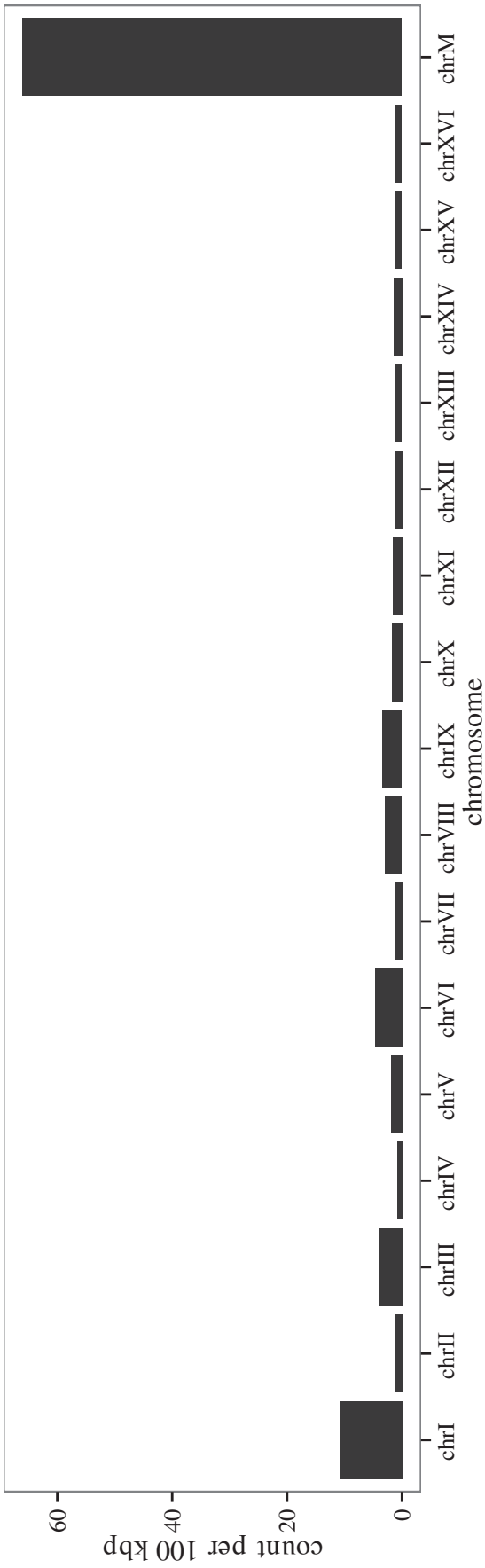
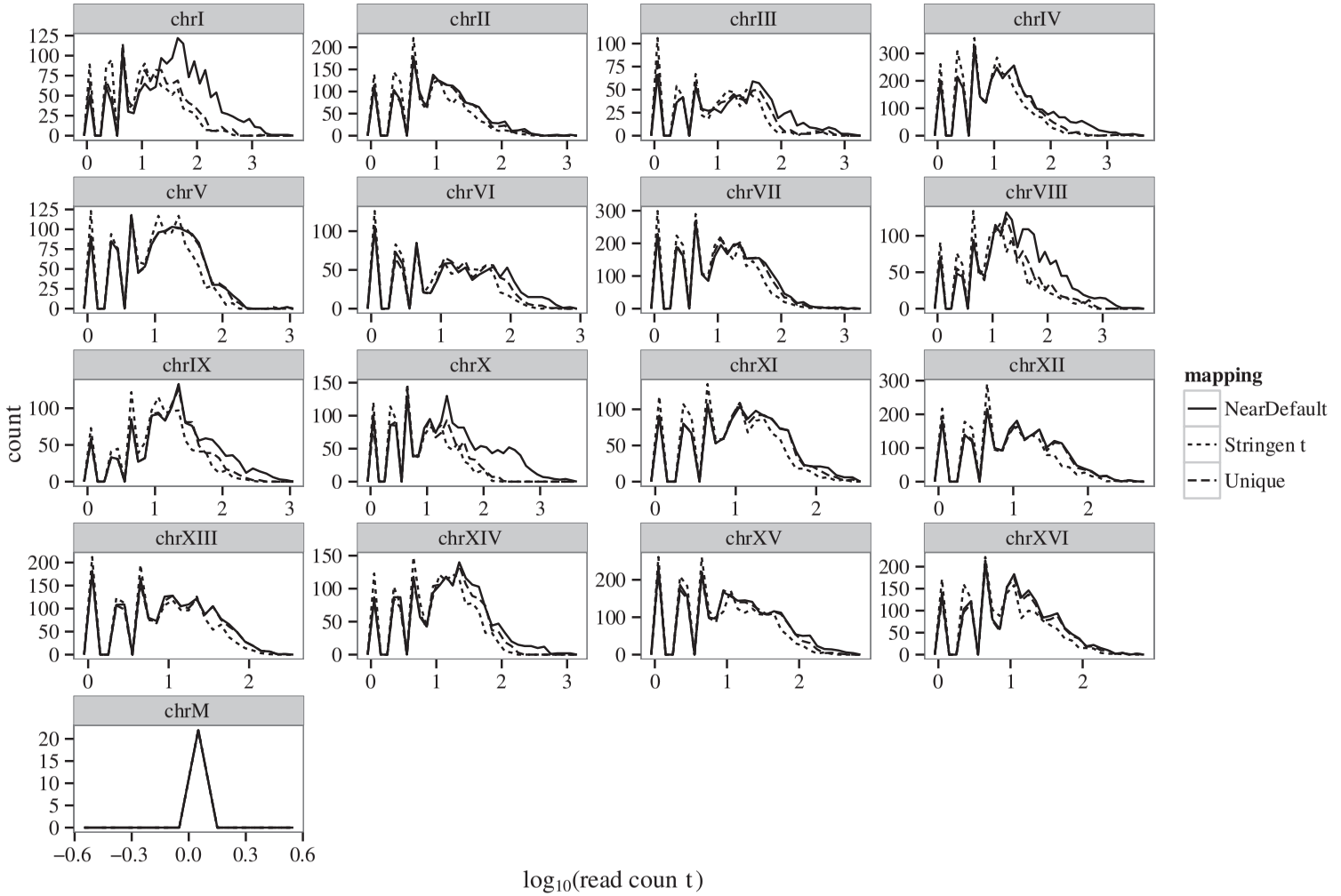


Figure 4.5: Number of un-annotated region open reading farmes per 100 kbp, calculated per chromosome.

The distribution of read counts amongst the UAR ORFs per chromosome are depicted in 4.6 for the three RNA-seq mapping methods. Across all but the mitochondrial chromosome, the majority of the distributions of read counts lie within 10-100 read counts. The asymetry in the distributions highlights that most UAR ORFs are only expressed at low levels. Distributions for the Near-Default, Unique, and Stringent mapping methods are relatively similar and overlap relatively well, with chromosome I having the largest difference between the Near-Default compared with Unique and Stringent mappings.

Figure 4.6: Distributions of read counts amongst the UAR ORFs per chromosome for the Near-Default, Unique, and Stringent mapping methods.



4.5.2 Searching for Un-Annotated Region Open Reading Frames

UAR ORF Minimum Length of 6 Amino Acids

The proteomics database that was searched against contained:

- 5,887 translated SGD gene sequences
- 5,887 reversed SGD gene sequences
- 10,794 translated UAR ORF sequences
- 10,794 reversed UAR ORF sequences

The 5,887 translated SGD gene sequences refer to the set of translations of all systematically named ORFs from SGD, except dubious ORFs and pseudogenes. The 1% FDRs are listed in Table 4.8.

The 1% FDR thresholds for protein-level searches for the 6-amino acid database are listed in Table 4.6. Across the 3 replicates, about 67-69% of the 5,887 protein-coding SGD genes were identified above the FDR threshold, and less than 1% of reversed SGD sequences were identified above the threshold (Table 4.7). Since only 69% of the 6,717 SGD sequences (including those that are verified, uncharacterised, dubious, transposable elements, and pseudogenes) were detected with the proteomics data in the original study (Tyagi and Pedrioli, 2015), this study's rate of identification of protein-coding SGD genes may be artificially low. There is a set of SGD protein-coding sequences that is inherently undetectable by this set of proteomics data. In replicate 3, UAR ORF 6534 was identified above the 1% FDR threshold and had a probability value of 0.975. Although UAR ORF 6534 is seen in neither replicates 1 nor 2, it has a very high probability value from replicate 3, warranting

Table 4.6: The 1% false-discovery rate thresholds for protein-level searches for databases for proteomics searches.

Proteomics Search Database	rep1	rep2	rep3
6aa	0.4633	0.5376	0.6204
12aa	0.4643	0.5435	0.621
23aa	0.4648	0.4659	0.621
Top627	0.7088	0.6611	0.674

Table 4.7: Proteins identified in protein-level proteomics searches for un-annotated region open reading frames at least 6 amino acids long. Multiple sequences = proteins with peptide spectrum matches to any combination of SGD sequences, reversed SGD sequences, UAR ORF sequences, and reversed UAR ORF sequences from the database.

Category	rep1	rep2	rep3
SGD sequences	4071 (69.2%) - % of 5,887	4068 (69.1%)	3960 (67.3%)
Reversed SGD sequences	4 (0.068%)	4 (0.068%)	3 (0.051%)
UAR ORF sequences	0	0	1
Reversed UAR ORF sequences	0	0	0
Multiple sequences	47	46	38
Total sequences	4122	4118	4002

further investigation.

For replicate 1 (rep1), a total of 215,242 peptide spectum matches were to at least one sequence in the proteomics search database. Of those, 89.1% matched just one SGD sequence above the 1% FDR threshold, and 18 peptide spectrum matches were to just one reversed SGD sequence in Figure 4.9. For rep2 and rep3, there were similar rates of peptide spectum matches to SGD sequences and reversed SGD sequences. In replicate 3, a peptide spectrum match to UAR ORF 5644 was

Table 4.8: The 1% false-discovery rates thresholds for protein-level searches for databases for proteomics searches.

Proteomics Search Database	rep1	rep2	rep3
6aa	0.74	0.7606	0.7823
12aa	0.7397	0.7609	0.7821
23aa	0.7402	0.74	0.7822
Top627	0.809	0.7931	0.8283
Rep964_1178bp	0.7833	0.8043	0.802

identified above the 1% FDR, with a probability value of 0.814. In the RNA-seq mappings, the ORF had 5 reads for Near-Default and Unique, and 4 reads for Stringent mapped to the region. Since reads remained even in the Stringent mapping that required a 100% match, without another location in the genome the reads could match at equal quality, these pieces of evidence require further exploration. In replicate 3, there was one peptide spectrum match to UAR ORF 6534 at a probability value of 0.865. In the RNA-seq data, for all 3 mappings methods, there were 4 50-bp reads that mapped to this region. Most importantly, these 4 RNA-seq reads remained even in the Stringent mapping that required a perfect mapping, providing evidence toward transcription of the region.

Table 4.9: Peptide spectrum matches from the proteomics search for un-annotated region open reading frames at least 6 amino acids long. Multiple sequences = peptide spectrum matches for any combination of SGD sequences, reversed SGD sequences, UAR ORF sequences, and reversed UAR ORF sequences from the database.

Category	rep1	rep2	rep3
SGD sequences	191792 (89.1%) - % of total peptide spectrum matches	219065 (89.7%)	150496 (89.5%)
Reversed SGD sequences	18 (0.0084%)	29 (0.012%)	12 (0.0071%)
UAR ORF sequences	0	1	1
Reversed UAR ORF sequences	1	1	0
Multiple sequences	23431	25213	17605
Total peptide spectrum matches	215242	244309	168114

Table 4.10: Peptide spectrum matches from the proteomics search for un-annotated region open reading frames at least 12 amino acids long. Multiple sequences = peptide spectrum matches for any combination of SGD sequences, reversed SGD sequences, UAR ORF sequences, and reversed UAR ORF sequences from the database.

Category	rep1	rep2	rep3
SGD sequences	191790 (89.1%)	219056 (89.7%)	150514 (89.5%)
Reversed SGD sequences	17 (0.0079%)	29 (0.012%)	12 (0.0071%)
UAR ORF sequences	0	1	1
Reversed UAR ORF sequences	1	1	0
Multiple sequences	23434	25214	17616

UAR ORF Minimum Length of 12 Amino Acids

The proteomics database that was searched against contained:

- 5,887 translated SGD gene sequences
- 5,887 reversed SGD gene sequences
- 6,643 translated UAR ORF sequences
- 6,643 reversed UAR ORF sequences

Total numbers of peptide spectrum matches above the 1% FDR thresholds are listed in Table 4.10. About 89% of all peptide spectrum matches were to only one SGD sequence, and less than 1% of all peptide spectrum matches were to reversed SGD sequences for each of the 3 replicates above their respective 1% FDR thresholds. Similarly, a peptide spectrum match in replicate 2 was to UAR ORF 5644, and a peptide spectrum match in replicate 3 was to UAR ORF 6534 above their respective 1% FDRs.

For protein-level searches, about 67-71% of the 5,887 SGD sequences from the databases were identified above the 1% FDR thresholds in Table 4.11. Less than 1% of the 5,887 reversed SGD sequences were identified. Much like the 6-Amino Acid database search, UAR ORFF 6534 was once again identified and had a probability

Table 4.11: Proteins identified in protein-level proteomics searches for un-annotated region open reading frames at least 12 amino acids long. Multiple sequences = proteins with peptide spectrum matches to any combination of SGD sequences, reversed SGD sequences, UAR ORF sequences, and reversed UAR ORF sequences from the database.

Category	rep1	rep2	rep3
SGD sequences	4070 (69.1%)	4191 (71.2%)	3960 (67.3%)
Reversed SGD sequences	4 (0.068%)	13 (0.22%)	3 (0.051%)
UAR ORF sequences	0	0	1
Reversed UAR ORF sequences	0	0	0
Multiple sequences	47	43	38
Total sequences	4121	4247	4002

value of 0.975 in replicate 3.

UAR ORF Minimum Length of 23 Amino Acids

The proteomics database that was searched against contained:

- 5,887 translated SGD gene sequences
- 5,887 reversed SGD gene sequences
- 2,898 translated UAR ORF sequences
- 2,898 reversed UAR ORF sequences

Regarding peptide-level searches, all 3 replicates had around 89% of the 5,887 known SGD sequences identified above the 1% FDR in Table 4.12. Again, less than 1% of the reversed SGD sequences were found above the 1% FDR. In replicate 3, there was a peptide spectrum match to UAR ORF 6534 above the 1% FDR; however, there were no peptide spectrum matches in replicate 2 to any UAR ORFs, unlike for 6-Amino Acid and 12-Amino Acid databases. In replicate 2, there was a peptide spectrum match to UAR ORF 5644, but the probability value was 0.0017631, well below the 1% FDR.

Table 4.12: Peptide spectrum matches from the proteomics search for un-annotated region open reading frames at least 23 amino acids long. Multiple sequences = peptide spectrum matches for any combination of SGD sequences, reversed SGD sequences, UAR ORF sequences, and reversed UAR ORF sequences from the database.

Category	rep1	rep2	rep3
SGD sequences	191895 (89.1%)	191888 (89.1%)	150602 (89.5%)
Reversed SGD sequences	18 (0.0084%)	18 (0.0084%)	13 (0.0077%)
UAR ORF sequences	0	0	1
Reversed UAR ORF sequences	1	1	0
Multiple sequences	23443	23447	17625
Total peptide spectrum matches	215357	215354	168241

Table 4.13: Proteins identified in protein-level proteomics searches for un-annotated region open reading frames at least 23 amino acids long. Multiple sequences = proteins with peptide spectrum matches to any combination of SGD sequences, reversed SGD sequences, UAR ORF sequences, and reversed UAR ORF sequences from the database.

Category	rep1	rep2	rep3
SGD sequences	4071 (69.2%)	4068 (69.1%)	3964 (67.3%)
Reversed SGD sequences	4 (0.068%)	4 (0.068%)	3 (0.051%)
UAR ORF sequences	0	0	1
Reversed UAR ORF sequences	0	0	0
Multiple sequences	48	46	38
Total sequences	4123	4118	4006

For protein-level proteomics searches, 67-69% of all SGD sequences were identified and less than 1% of reversed SGD sequences were detected at probability values above the 1% FDR thresholds for all 3 replicates 4.13. Like the 6-Amino Acid and 12-Amino Acid database searches, UAR ORF 5634 was mapped at a probability of 0.860262.

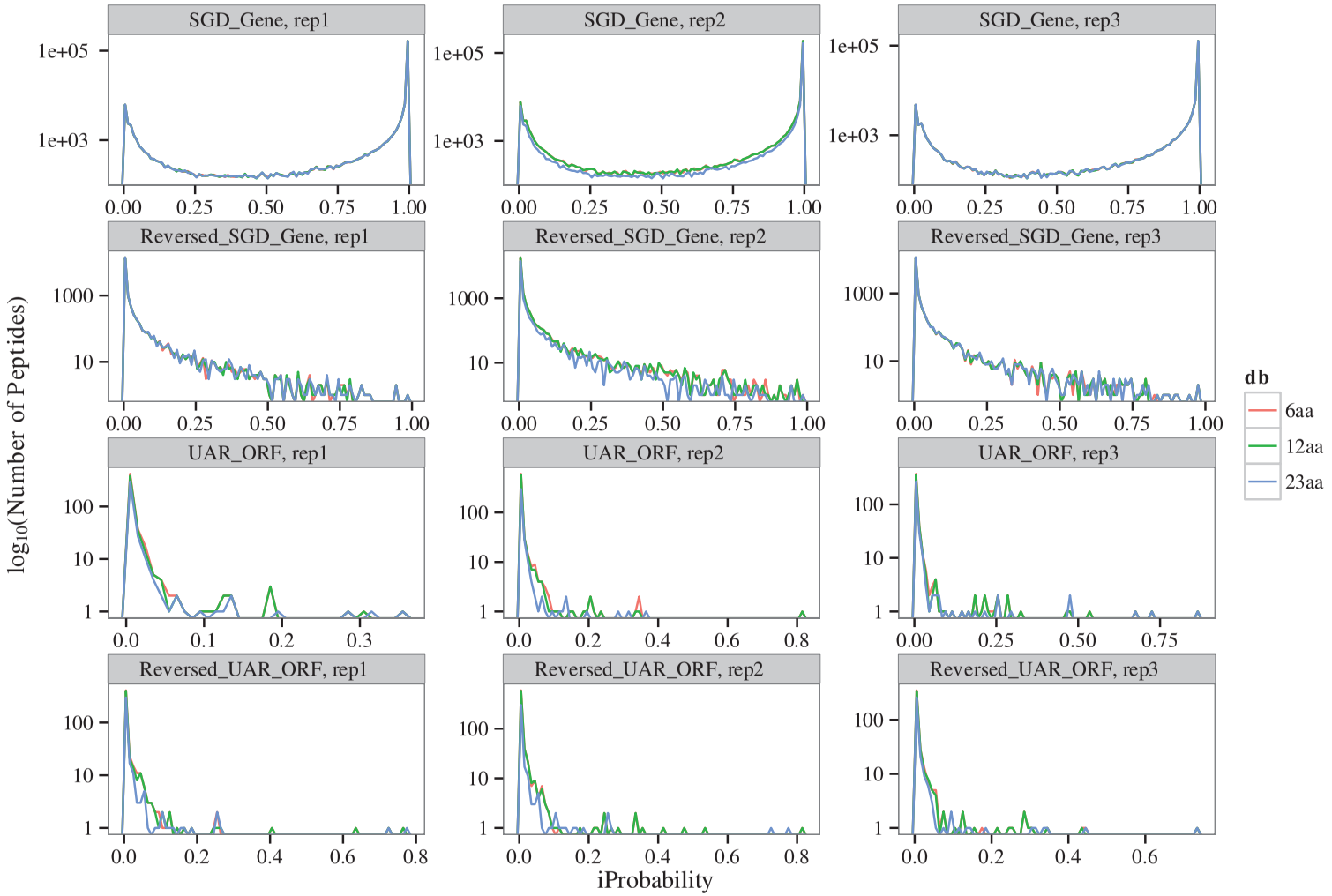
4.5.3 Comparison of the 6-, 12-, and 23-Amino Acid Databases

Although there were slight variations in the actual values denoting the 1% FDR thresholds for peptide-level proteomics searches for the 3 different databases, there were no solid trends. For example, for replicate 1, the 6-Amino Acid database had

a value of 0.7400, which was the value decreased for the 12-Amino Acid one, then increased for the 23-Amino Acid database. For replicate 2, the value increased then decreased. For replicate 3, the FDR threshold decreased then slightly increased. Similar behavior is evident in p-values for the 1% FDRs for protein-level searches. Therefore, changing the size of databases (adding from 10,794 UAR ORF sequences for the 6-Amino Acid database, 6,643 for 12-Amino Acid, or 2,898 for 23-Amino Acids) did not appear to have a profound effect on the 1% FDR thresholds or the percentage of SGD sequences identified on the protein level.

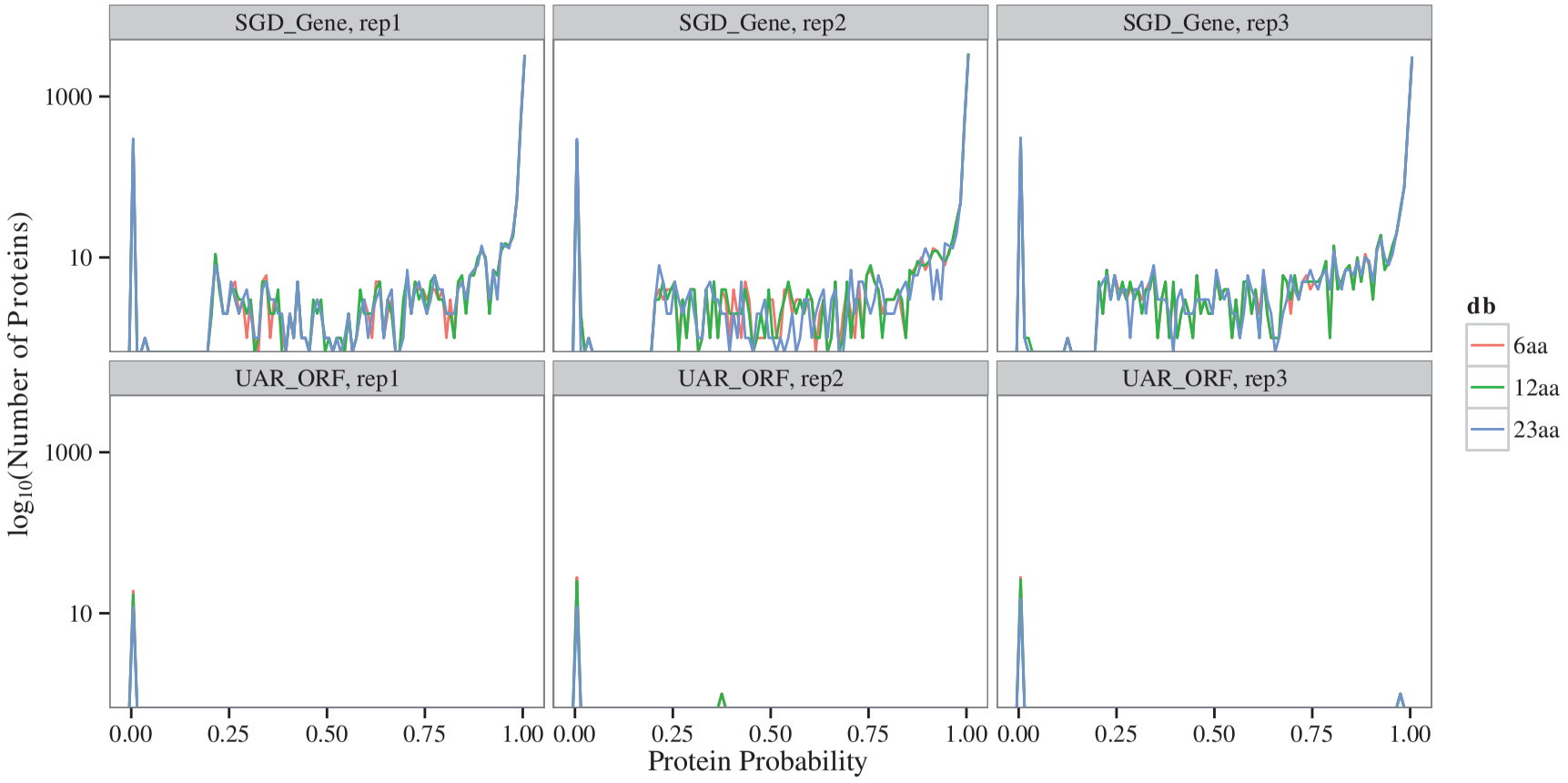
For peptide-level searches Figure 4.7 shows the distributions of iProbability values for all sequence types across the 3 databases for each replicate. Overall, there is great overlap amongst the histograms for all 3 databases.

Figure 4.7: Distributions of iProbability values for SGD protein-coding genes, un-annotated open reading frames, and their respective reversed sequences for the 6-aa, 12-aa, and 23-aa proteomics databases.



For protein-level searches, Figure 4.8 gives distributions of Protein Probability values for the SGD genes and UAR ORF sequences for the 3 databases for each of the replicates. Again, overall, there is significant overlap in the distributions across the 6-, 12-, and 23-Amino Acid databases.

Figure 4.8: Distributions of Protein Probability values for SGD protein-coding genes and un-annotated region open reading frames for the 6-aa, 12-aa, and 23-aa proteomics databases.



4.6 Curation of a Database of Protein-Coding Genes with High mRNA Expression

Exploiting the fact that the *Saccharomyces cerevisiae* genome is one of the best curated eukaryotic genomes, the set of known RNA transcripts/genes can serve as a set of candidates that are likely to be detected by proteomics. Other peptide sequences may serve as a set of the least likely candidates to be detected by proteomics. Treating these two groups almost as if they were true positives and true negatives, respectively, values that may resemble sensitivity and specificity can be calculated to further characterise these methods (RNA-Seq and proteomics -- proteogenomics) in terms of database size, types of sequences, different sets of gene models given to pipelines, etc. This new knowledge will inform and direct future studies aimed at discovering new peptides or protein-coding genes in less well-curated organisms in using typical proteogenomics approaches.

Two databases of comparable size, in terms of total number of SGD protein-coding sequences, were constructed to simulate a set of true positives to search the proteomics data against. One was constructed based solely on the number of Near-Default RNA-sequencing reads that mapped to each sequence; whereas the other was constructed based on the median length of 1,071 protein-coding SGD sequences.

4.6.1 Top 627 SGD Sequences with Highest Number of Reads

The proteomics database (Top627) that was searched against contained:

- 627 SGD gene sequences
- 627 reversed SGD gene sequences

All 6,603 total SGD gene sequences were sorted in descending order according to the number of RNA-seq reads in the Near-Default mapping. The top 10% consisted of 660 SGD sequences, which were then mapped to the 5,887 translated systematically named ORFs from SGD, leaving 627 SGD gene sequences that were included in that set.

The distribution of lengths for the 627 select SGD sequences was very different from that of the entire set of 5,887 translated SGD sequences (Figure 4.9). The median length of the 5,887 sequences was 1,071 bp, whereas that that for the 627 top sequences was 413,000 bp.

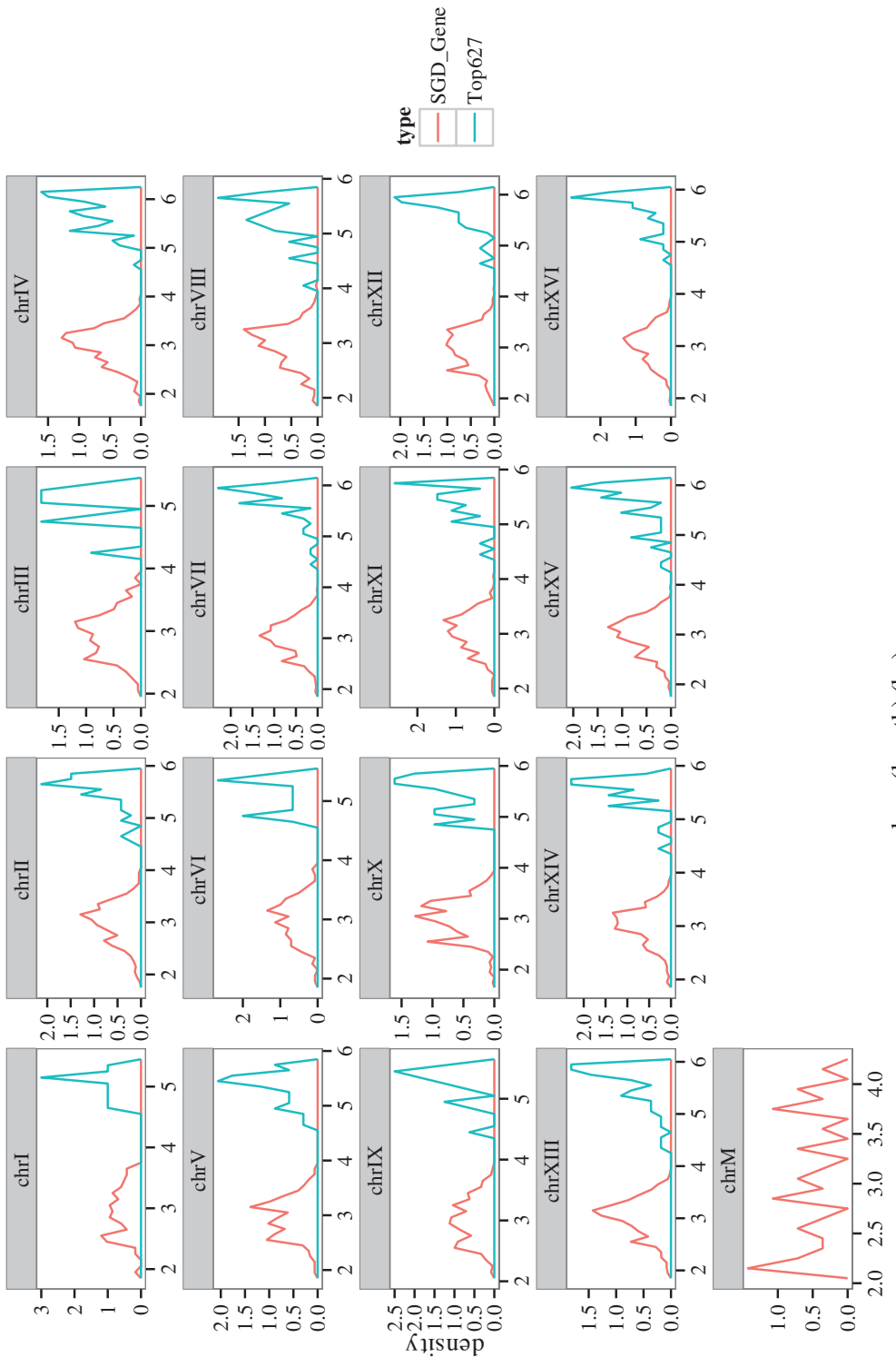


Figure 4.9: Probability distributions of the lengths of SGD protein-coding genes and the Top 627 sequences.

Table 4.14: Peptides identified during the proteomics search for the Top 627 SGD protein-coding sequences. Multiple sequences = proteins with peptide spectrum matches to any combination of SGD sequences and reversed SGD sequences from the database.

Category	rep1	rep2	rep3
SGD sequences	98665 (82.7%)	113439 (82.8%)	80392 (82.8%)
Reversed SGD sequences	75 (0.063%)	94 (0.069%)	40 (0.41%)
Multiple sequences	20595	23423	16603
Total peptides	119335	136956	97035

Table 4.15: Proteins identified in protein-level proteomics searches for the Top 627 SGD protein-coding sequences. Multiple sequences = proteins with peptide spectrum matches to any combination of SGD sequences and reversed SGD sequences from the database.

Category	rep1	rep2	rep3
SGD sequences	509 (81.2% of whole db)	512 (81.7%)	511 (81.5%)
Reversed SGD sequences	31 (4.9%)	31 (4.9%)	26 (4.1%)
Multiple sequences	36	33	32
Total sequences	576	576	569

In the peptide-level searches, about 82% of all peptides were assigned to SGD sequences at values above the 1% FDR thresholds for all 3 replicates (Table 4.14). Less than 1% of reversed SGD sequences were detected above the 1% FDR thresholds for each replicate.

For the protein-level searches, about 81% of the 627 SGD sequences were identified above the 1% FDR threshold for each replicate (Table 4.15). However, almost 5% of the reversed SGD sequences were detected at probabilities above the 1% FDR thresholds, higher than any of the previous databases mentioned already.

4.6.2 Representative SGD Sequences at 964-1178 bp in Length

The proteomics database (Rep964_1178bp) that was searched against contained:

- 626 SGD gene sequences
- 626 reversed SGD gene sequences

The median length of all 6,603 SGD gene sequences was found to be 1,071 bp. To have a fair comparison of this database with Top627, the two databases should be of the same size. Adding and subtracting 10% of 1,071 bp gave a range of 964 bp to 1,178 bp. All 6,603 SGD genes with lengths in this range were selected for the Rep964_1178bp, which was comprised of 626 select SGD genes.

Unlike Top627, the lengths of the sequences in Rep964_1178bp reflected the probability distribution of the entire set of SGD genes much better in (Figure 4.10). In this case, for most chromosomes, the majorities of both probability distributions overlapped. In addition, the RNA-seq read count distribution for Rep964_1178bp sequences also reflected patterns seen for the entire set of SGD genes (Figure 4.11). There is a large spread of low numbers of sequences with low read counts, a large peak, and a sharp tapering to low numbers of sequences with very high read counts. Again, the distributions for the most of chromosome overlap, especially where the majority of the distributions lie. As the Top627 sequences were chosen by taking the highest read counts, it would be expected that probability distributions for that set would be skewed to the right, not reflecting distributions of the entire set of SGD genes in Figure 4.9.

Figure 4.10: Probability distributions of the lengths of SGD protein-coding genes and Representative sequences that are between 964 and 1178 bp.

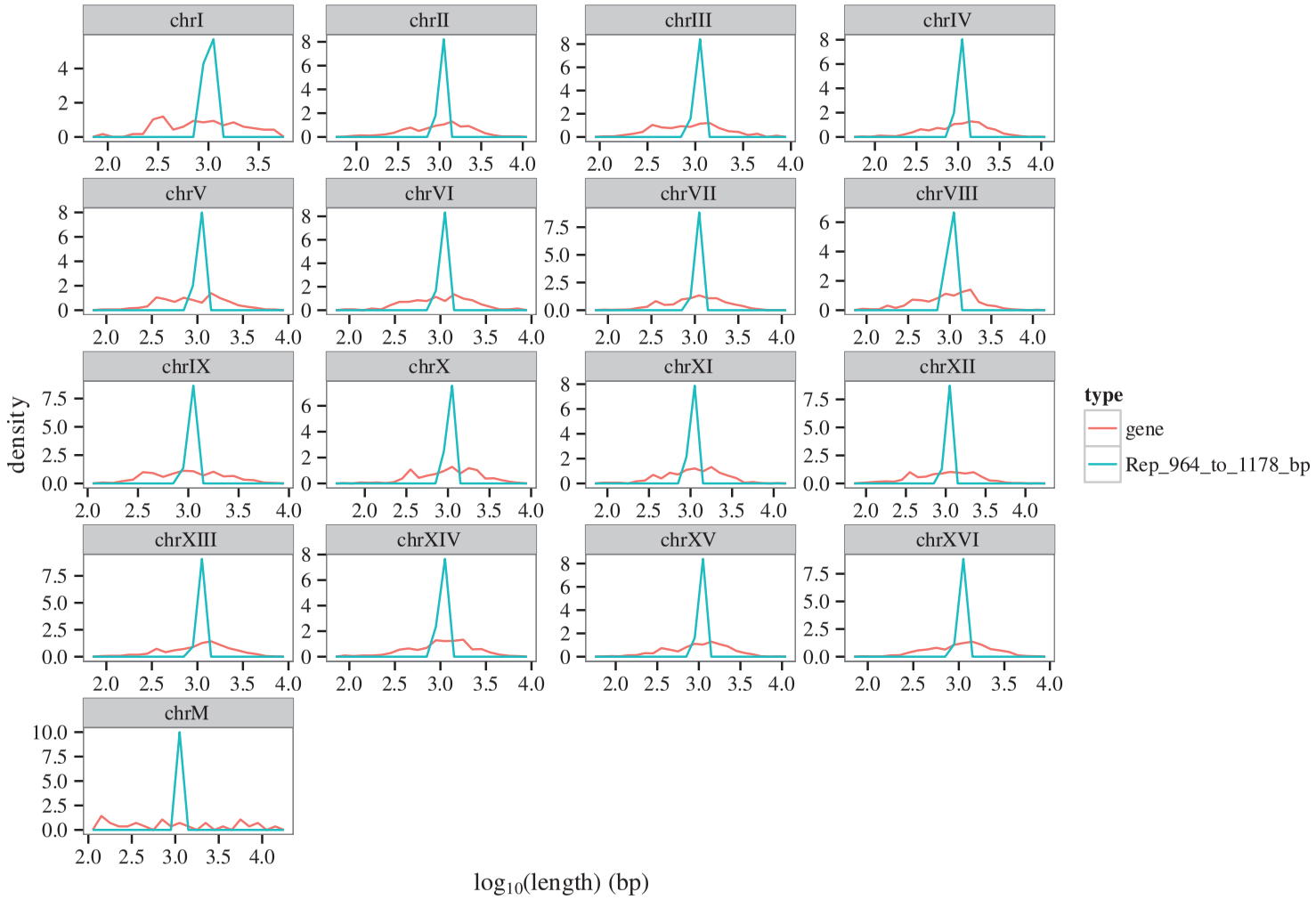


Figure 4.11: RNA-seq read count distribution for SGD protein-coding genes and Representative sequences that are between 964 and 1178 bp, per chromosome.

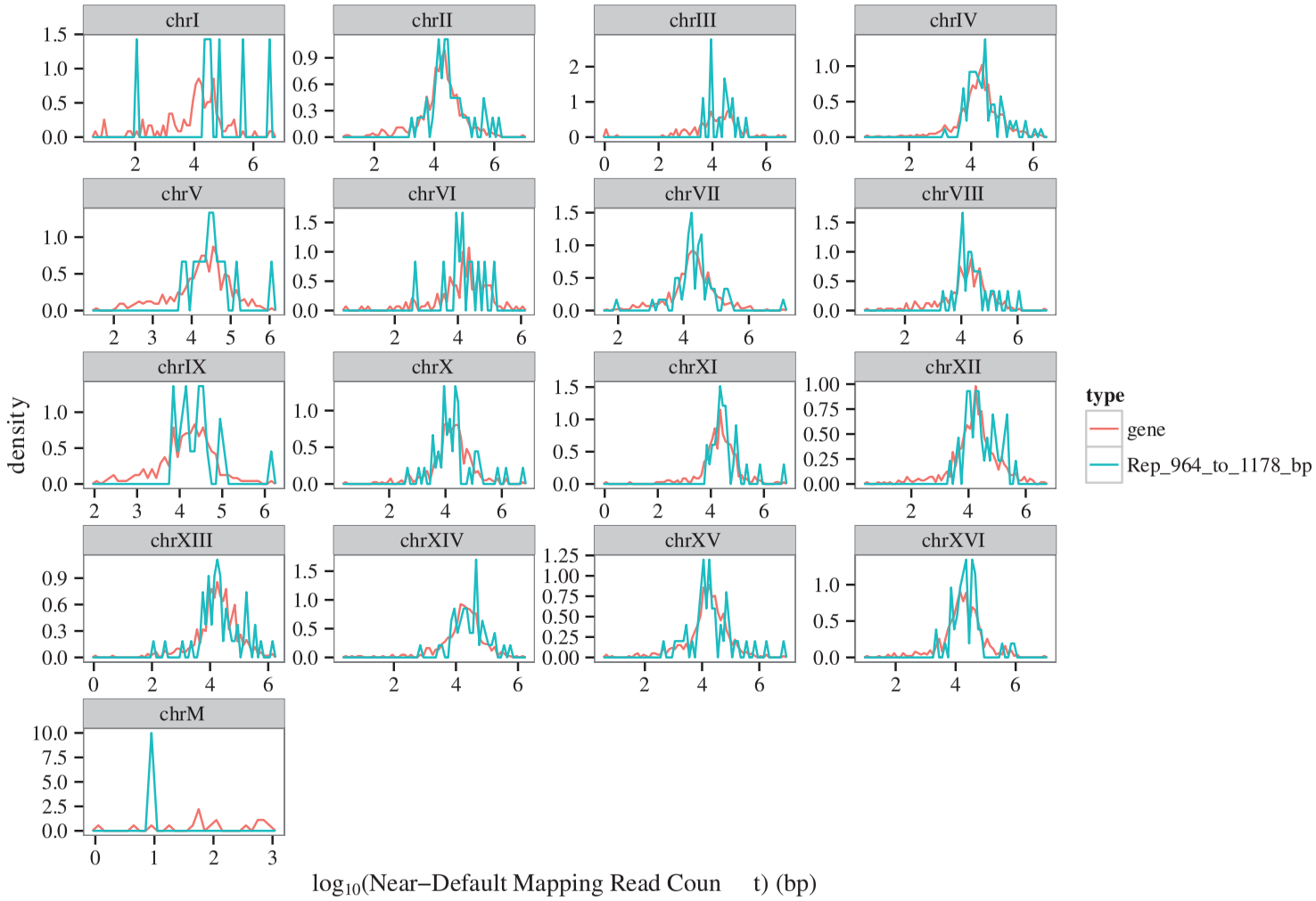


Table 4.16: Peptide spectrum matches from the proteomics search for the Representative SGD protein-coding sequences between 964 and 1178 bp. Multiple sequences = peptide spectrum matches for any combination of SGD sequences and reversed SGD sequences from the database.

Category	rep1	rep2	rep3
SGD sequences	28493 (87.1%)	33029 (87.3%)	23526 (87.1%)
Reversed SGD sequences	6 (0.018%)	9 (0.024%)	2 (0.0074%)
Multiple sequences	4226	4811	3478
Total peptide spectrum matches	32725	37849	27006

For peptide-level searches, over 87% of all peptide spectrum matches were to one SGD sequence, and less than 1% of all peptide spectrum matches were to a reversed SGD sequence for all replicates (Table 4.16).

For protein-level searches, 70-73% of sequences within the Rep964_1178bp were detected at probabilities above the 1% FDR thresholds for each of the 3 replicates (Table 4.17). Less than 1% of reversed SGD sequences were mapped above the 1% FDRs for all replicates.

4.6.3 Comparison of Top627 and Rep964_1178bp Databases

On the peptide-level, many more peptide spectrum matches were produced with the Top627 database than the Rep964_1178bp. This may be due to the proteins being highly expressed and thus more likely to be present in the sample and readily detected (Figure 4.12). Although there were many more peptides in Rep964_1178bp, the distributions followed the same behavior for both databases. Interestingly, about 5% more of the total peptide spectrum matches from Rep964_1178bp were to only one SGD sequence compared with the Top627 database (about 87% compared with about 82%). Regarding the 1% FDR thresholds for peptide-level searches, the values in Table 4.8 slightly varied within replicates. The differences between values for

Table 4.17: Proteins identified in protein-level proteomics searches for the Representative SGD protein-coding sequences between 964 and 1178 bp. Multiple sequences = proteins with peptide spectrum atches to any combination of SGD sequences and reversed SGD sequences from the database.

Category	rep1	rep2	rep3
SGD sequences	457 (73.1%)	461 (73.8%)	443 (70.9%)
Reversed SGD sequences	4 (0.64%)	2 (0.32%)	1 (0.16%)
Multiple sequences	4	3	5
Total sequences	465	466	449

Top627 and Rep964_1178bp were smaller than differences amongst the thresholds of

6-, 12-, or 23-Amino Acid databases.

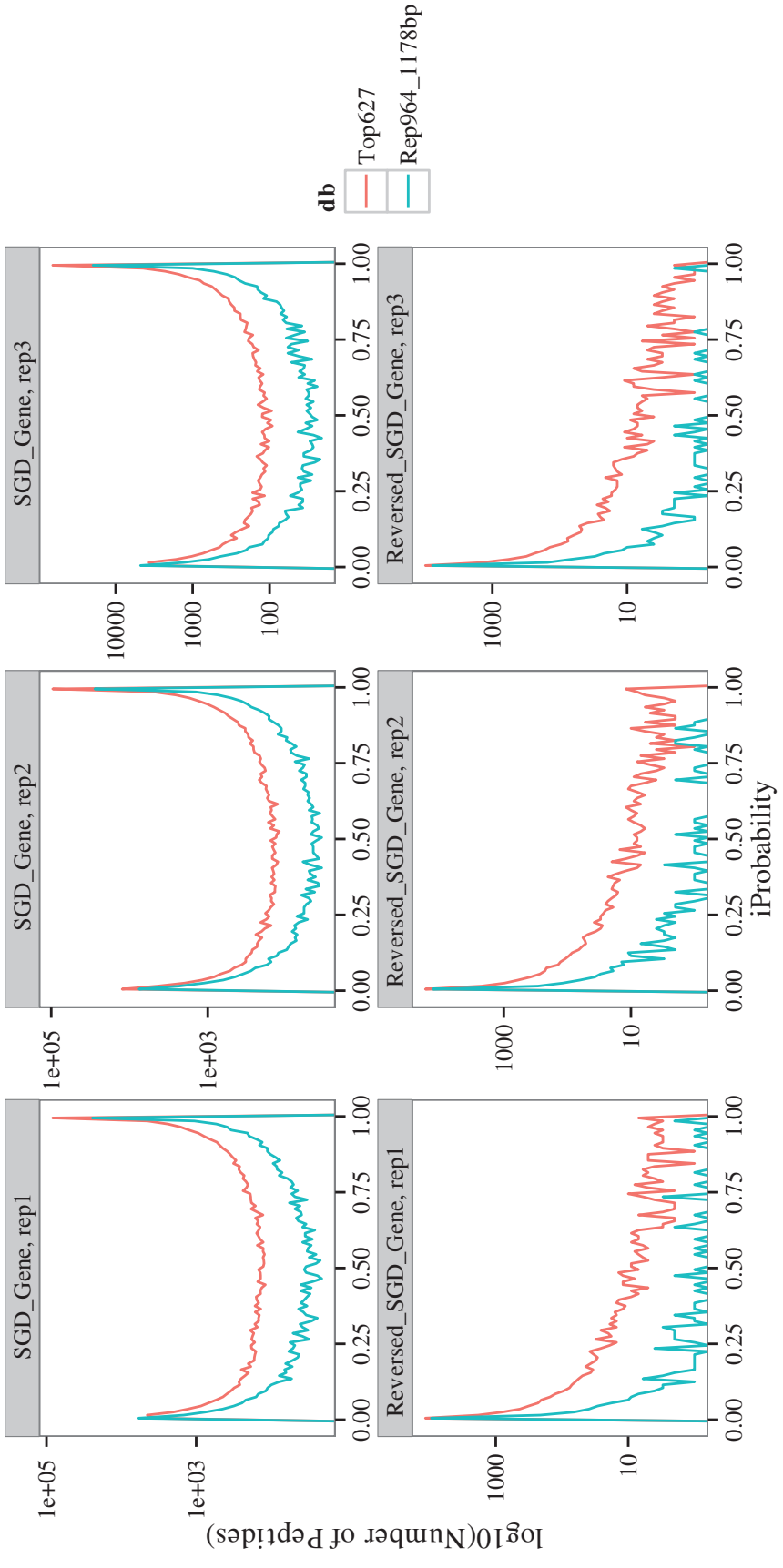


Figure 4.12: Distributions of the number of peptide spectrum matches across iProbability values for the Top 627 sequences and Representative sequences between 964 and 1178 bp.

For protein-level searches, about 81% of SGD sequences were mapped for Top627, whereas 70-73% were for Rep964_1178bp above 1% FDR thresholds. However, the rates of false positives (mapped reversed SGD sequences) for Top627 were higher at around 5%, and in fact, are the highest of all the databases searched in this project. Comparing these two databases with the 6-, 12-, and 23-Amino Acid series of databases, the Rep964_1178bp search results are more similar. The 6-, 12-, and 23-Amino Acid databases yielded about 67-71% of SGD sequences mapped, closest to the 70-73% for Rep964_1178bp. The false positive rates of less than 1% were also common between the length-filtered databases and Rep964_1178bp, whereas Top627 had about 5% of reversed SGD sequences detected. Additionally, the distributions of Protein Probability values are relatively similar between the two databases 4.13.

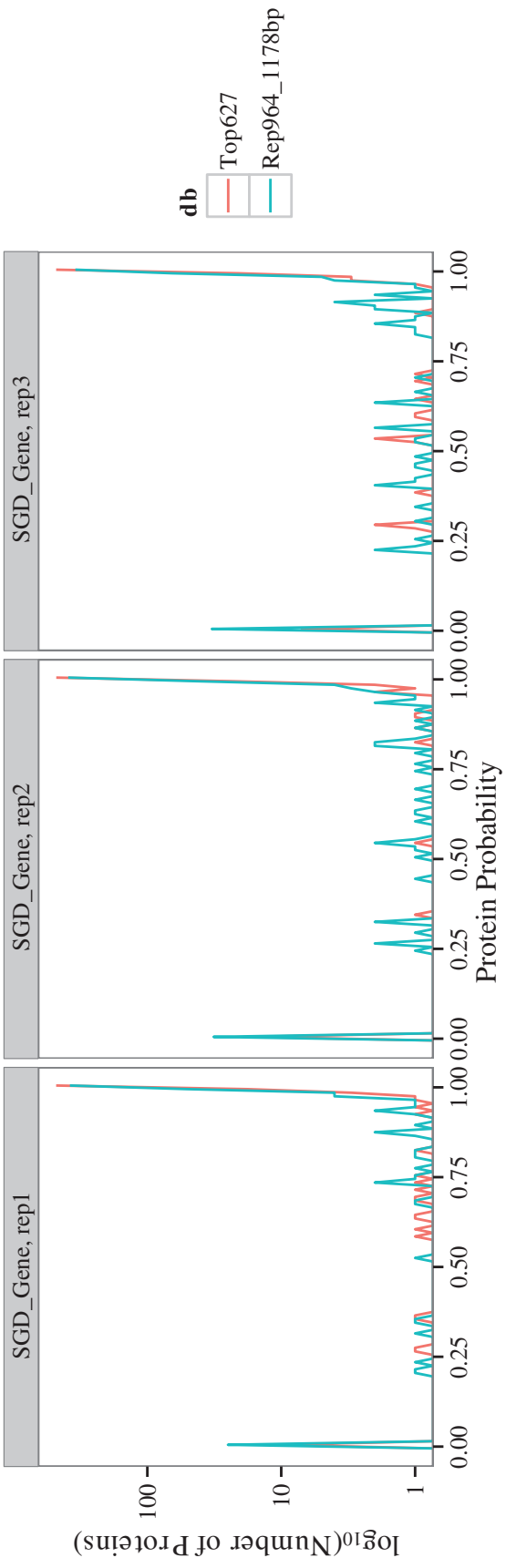


Figure 4.13: Distributions of the number of peptide spectrum matches across values of Protein Probability for proteomics searches against the Top 627 and Representative 964-1178 bp databases.

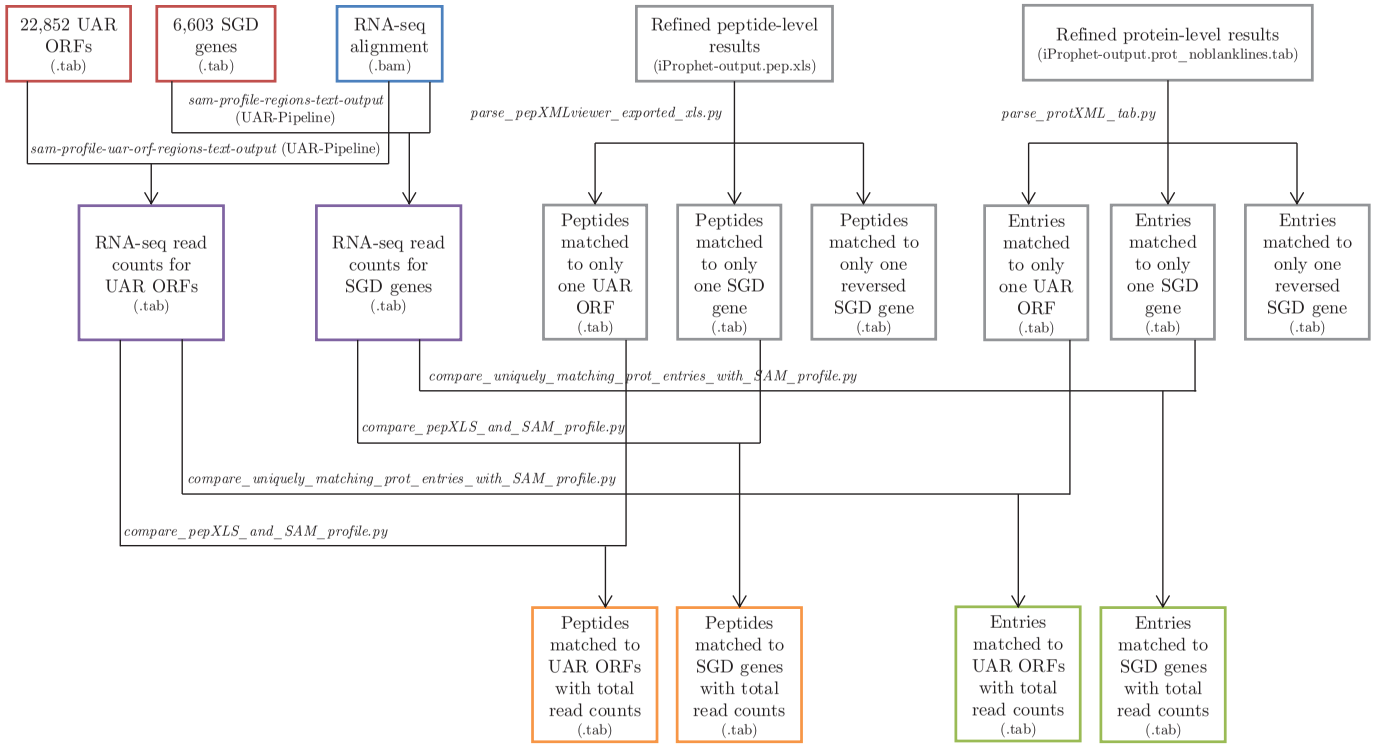
Overall, the Rep964_1178bp and Top627 databases can be used to compare against proteomics search results with new or unknown target sequences (e.g. containing UAR ORFs). The Rep964_1178bp database has the advantage of better representation of sequence length, while the Top627 database has the advantage of including highly-expressed genes. Both methods would be applicable in searching the genomes of less well-annotated organisms.

4.7 RNA-seq and Proteomics Analysis

4.7.1 Work Flow

After generating the results from the proteomics analysis, the question of which UAR ORFs with RNA-seq read counts also had a high probability of being expressed on the peptide level remained. To investigate, it was necessary to bring together output from the UAR-Pipeline (Section 2.5) with output from the proteomics analysis described in earlier sections of this chapter. A schematic diagram in Figure 4.14 illustrates how the task was achieved.

Figure 4.14: Schematic diagram of how RNA-seq read count output from the UAR-Pipeline (Section 2.5) and peptide spectrum matches and protein identifications from the proteomics pipeline are combined. Scripts are italicised.



With the UAR-Pipeline, RNA-seq read counts from the Near-Default alignment were found for the 22,852 UAR ORFs of interest by the module *sam-profile-uar-orf-regions-text-outout*. Read counts were also found for the 6,603 known SGD genes in a similar fashion, but with the module *sam-profile-regions-text-output*. These steps produced one .tab file of all UAR ORFs with their respective read counts, and one .tab file for the SGD genes also with read counts.

It was observed that on the peptide-level results, some individual identifications would be mapped to multiple sequences from the sequence database. For example, one identification mapped to both a known SGD gene sequence and an UAR ORF. To reduce the ambiguity of which sequences peptide spectrum matches were matching to, only those that were mapped to a single sequence were considered in subsequent analysis. To extract peptides that uniquely matched single sequences, the script *parse_pepXMLviewer_exported_xls.py* was written. The script returned three sets of results - one set for peptide spectrum matches for SGD genes, for reversed SGD genes, and for UAR ORFs.

Similarly, some protein-level results contained identifications to multiple sequences as well. Likewise, only protein identifications mapped to a single sequence were considered. Here, the *parser_protXML_tab.py* script was written to find protein identifications uniquely mapping UAR ORFs, SGD genes, or reversed SGD genes. Much like the parsing of peptide spectrum match results, this script returned three sets of results, one for each of the three categories of sequences. Moreover, to evaluate the quality of the protein-level identifications, for each sequence category, the non-redundant total number of sequences detected across all three replicates was determined. The number of SGD genes detected, 4,833 in total, was comparable to

that found in Tyagi and Pedrioli (2015), confirming the quality of the proteomics searches performed in this study.

After determining which peptide spectrum matches and protein identifications uniquely mapped to a single database sequence, RNA-seq read counts from the Near-Default mapping were found for sequences matched in these identifications. For peptide spectrum matches, this task was accomplished with the *compare_pepXLS_and_SAM_profile.py* script, which used the sequence IDs to cross-match peptide spectrum match identifications with read counts. The comparison resulted in a .tab file containing read counts for peptide spectrum match identifications for UAR ORFs and another .tab file for SGD genes. Similarly, this was done for protein identifications with the script *compare_uniquely_matching_prot_entries_with_SAM_profile.py*.

4.8 Jackknife Testing

To test the sensitivity and specificity of the Un-Annotated Region (UAR) and Proteomics Pipeline, known genes were removed from the original genome annotations in an attempt to be discovered by the naive Pipeline. The set of 6,600 known genes from the *Saccharomyces* Genome Database (SGD) were randomized and divided into 20 groups of 330 genes each. The jackknife set-up is such that there are 20 pipeline runs, and in each run only 19 groups of 330 genes are included in the known SGD genes database, leaving a different group of 330 genes to be 'discovered.'

For each group, annotations for the 330 genes were removed from the .gff3 file. The UAR Pipeline was run on the modified .gff3 file to determine the set of UARs and the open reading frames (ORFs) within the UARs. A FASTA database was

then constructed using the following:

- sequences of the remaining 6,270 SGD genes
- reversed sequences of the remaining 6,270 SGD genes
- sequences of the UAR ORFs at least 17 amino acids in length
- reversed sequences of the UAR ORFs at least 17 amino acids in length

The Proteomics Pipeline was then run to search the proteomics data against the FASTA database. Across all 20 groups, a total of 111 non-redundant UAR ORFs belonging to 110 non-redundant masked SGD genes were detected by the Pipeline (Table 4.18). In Group 12, there were 2 different UAR ORFs belonging to the masked gene YIL156W, which accounts for the discrepancy in numbers. The 110 non-redundant masked SGD genes were searched against SGD's YeastMine database (Balakrishnan et al., 2012) for a summary of each gene (Appendix D). Although UAR ORFs mapping to Masked SGD genes are listed in Table 4.18, only 1.7% of the 6,600 total Masked SGD genes were detected.

To assess the performance of the Proteomics Pipeline in discovering 'new' proteins, the quality of the raw proteomics data must be accounted for. In the original study Tyagi and Pedrioli (2015), only proteins with at least two high-confidence peptides detected across label-switched samples were used to calculate relative protein abundances. From the filtering, a total of 3,614 SGD proteins remained.

Of the 3,614 SGD proteins, 3,613 were included in the entire 6,600 set of SGD proteins considered in the jackknife testing and thus used to assess performance. For each jackknife group, the values below were determined to calculate sensitivity and specificity.

Table 4.18: For each Jackknife Group, the number of non-redundant Un-Annotated Region Open Reading Frames and number of non-redundant Masked SGD genes with a score over the 1% FDR threshold are listed.

Group	Number of UAR_ORFs with a score over the 1% FDR threshold	Number of Masked SGD genes with a score over the 1% FDR threshold
1	7	7
2	6	6
3	9	9
4	3	3
5	3	3
6	7	7
7	6	6
8	4	4
9	6	6
10	5	5
11	2	2
12	7	6
13	10	10
14	6	6
15	4	4
16	4	4
17	4	4
18	6	6
19	7	7
20	5	5
total	111	110

- FSSP (Filtered Selected SGD Proteins) = number of proteins in common between the 3,613 filtered proteins and the 330 selected SGD proteins to test per group
- TP (True Positives) = FSSP that were detected below the 1% FDR (per rep)
- FP (False Positives) = reversed sequences of FSSP that were detected below the 1% FDR (per rep)
- FN (False Negatives) = FSSP that were not detected below the 1% FDR (per rep)
- TN (True Negatives) = reversed sequences of FSSP that were not detected below the 1% FDR (per rep)

Values for all groups and all 3 biological replicates for each group are listed in Appendix E. Values for Group 1 and Rep 1 are calculated below as an example. Across all reps and groups, the average sensitivity and specificity for discovering 'new' proteins were 1.6% and 98.6%, respectively. The low sensitivity may be attributed to the fact that only completely un-annotated regions were included downstream analysis. The following example illustrates expected challenges created by this requirement. GeneA (100 bp) and GeneB (200 bp) each consists of a single ORF. Ten percent of GeneA's sequence does not overlap with any other annotation, but the remaining 90% overlaps with GeneB. Only GeneA is selected to be in the group of 330 masked SGD genes. From the revised annotation file in which the 330 masked SGD genes have been excluded, the UAR Pipeline will detect only the 10% of GeneA's sequence that does not overlap with any other annotation. If GeneA

consists of only a single ORF, then the probability of being detected in the proteomics searches also depends on how highly expressed the mRNA is and whether the 10% portion of the sequence will not be enzymatically cleaved during the sample preparation. A potential improvement of the UAR Pipeline may invoke a method by which un-annotated regions are detected in the current manner while incorporating a method that searches both upstream and downstream of the UAR to determine if the ORF continues and/or whether other adjacent ORFs may be a continuation of the potential novel peptide/protein. Single or multiple gene modelers may further refine which sequences are more likely to be potential peptides/proteins by comparing the original UAR, the extended UAR that includes the entirety of the ORF, and combinations of the original UAR with any adjacent ORFs.

- FSSP = 176
- TP = 2
- FP = 1
- FN = 176-2 = 174
- TN = 176-1 = 174
- sensitivity = TP/(TP + FN) = 2/(2+174)*100% = 1.1%
- specificity = TN/(FP + TN) = 174/(1+174) = 99.4%

The same calculations were also performed for Filtered Unselected SGD Proteins (FUSP; number of proteins in common between the 3,613 filtered proteins and the 6,270 remaining unselected SGD proteins per group). Across all reps and groups, the average sensitivity and specificity for detecting FUSP were 97.0% and 99.8%, respectively.

Chapter 5

Discussion and Future Work

5.1 Introduction

In this chapter, results of the entire study are discussed in a larger context. The effectiveness of methods used in this study, such as the three stringency levels of RNA-seq read alignment, the categorising of genome annotations into two distinct groups, and the creation of an IGB Quickload Site, is explained. Future work to improve upon the progress made through this study is also suggested in this chapter, especially in terms of making these new resources publically available to the research community and experimentally verifying the two preliminary un-annotated region open reading frame targets that were found through both RNA-seq and proteomics analyses.

5.2 Discussion

5.2.1 RNA-seq Alignments

The three different alignment methods, Near-Default, Unique, and Stringent, were instrumental in examining the RNA-seq read mappings. Genomic regions where multi-mapping reads were present were denoted by a decrease in the number of reads in the Unique alignment when compared to the Near-Default. Yet another decrease in the number of reads in a specific genomic region in the Stringent mapping as compared to the Unique is an indication of mismatches, insertions, or deletions present in the reads that were absent in the Stringent mapping. By having these successive layers of restrictions on read alignment, multi-mapping reads and reads that do not match exactly to the genome could be detected. In addition, the successive stringencies provided a corresponding successive increase in confidence in the actual genomic locations where reads were originally transcribed from. With the Stringent Mapping, unless there were instrumental sequencing errors, the locations where reads aligned possess the greatest accuracy since there were no other genomic locations the reads could map to better (no multi-mapping reads), and the genomic locations matched the read 100% (no mismatches, insertions, or deletions).

5.2.2 Primary and Secondary Annotations

By categorising the genome annotations into two pools, Primary and Secondary Annotations, genes, such as snoRNAs and proteins, and genetic interactions, such as DNA double-strand break hotspots and histone binding sites, could be distinguished. In order to determine whether new genomic features existed throughout the study,

it was imperative to clearly define the locations of the currently known genomic features. However, genetic interactions, such as protein binding sites, should not be considered a Primary Annotation since the primary sequence at that location could also code for an undiscovered peptide, for instance. Therefore, by using only Primary Annotations to determine where the intergenic regions are located, genomic regions that have annotation regarding interactions but not whether any genes are coded by the primary sequence are not excluded in the search for new genomic features. By considering also Secondary Annotations, other possible indications of expression, such as transcription factor binding sites, can be analysed against potential new genomic features.

5.2.3 IGB Quickload Site

An Integrated Genome Browser Quickload Site was constructed throughout this study and can be made publically available through the Internet after publication. All of the RNA-seq data, genome annotations, and other information tracks, such as conservation, will be available for download and use. As the vast majority of the information contained within the Site has been derived from the *Saccharomyces* Genome Database, users of SGD would be able to view all of the published datasets hosted by SGD in the IGB Quickload Site against the RNA-seq dataset analysed in this study or any other RNA-seq dataset.

5.2.4 Preliminary Targets

The first preliminary targets at genomic region at chrXII: 489,949-490,404 had a high sequence identity with a known rRNA in the yeast genome. The region was

initially detected with TopHat2 RNA-seq alignments; however, after alignment of the same dataset with STAR, the vast majority of reads aligning to the region were absent. As STAR was used exclusively for the three alignment methods further investigations into how TopHat2 differs from STAR were not pursued.

The second preliminary target, chrI: 12,427-13,361, has sequence homology to flocculin proteins. However, due to the lack of the functional PA14 flocculin domain and the small number of Flocculin type 3 repeats in the sequence, any protein coded by this region would most likely be able to function as a flocculin. Therefore, the region is most likely a pseudogene.

The third preliminary target, chrV: 288,525-290,125 shared some sequence similarity with Cdc4 (cell division control protein 4). However, due to a premature stop codon, any protein produced from this region would be truncated and most likely non-functional. This target region is probably another pseudogene.

The potential pseudogenes could be submitted to the *Saccharomyces* Genome Database to help refine current annotations for the yeast genome.

5.2.5 6-, 12-, and 23-Amino Acid Proteomics Databases

The size of the proteomics databases searched against containing 6-, 12- or 23-amino acid UAR ORFs could have been reduced to include only sequences which have RNA-seq read aligned associated with them. A smaller database would not only decrease computing time but also increase sensitivity, as demonstrated in Blakeley et al. (2012)). The Blakeley et al. (2012) study recommends several methods to increase sensitivity by removing redundancy in nucleotide sequence databases, which could be evoked in future work:

- before calculating the FDR, select the most likely frame for each nucleotide sequence based on the number of peptide spectrum matches
- search nucleotide sequences against protein bases using BLASTX to select the most likely frame based on homology
- if no homologues are available for the nucleotide sequence, use amino acid composition or codon usage to select the most likely frame

Although the aforementioned strategies may reduce the size of the UAR ORF databases, one disadvantage may be excluding potential new peptide or protein sequences detectable by the proteomics data. For a single nucleotide sequence, if there were 2 or more frames that were likely to produce peptides but only the most likely was selected, then only one peptide, instead of multiple peptides, would be detected.

Perhaps one way to determine our set of true negatives is all of the potential ORFs that did not match in our proteomics searches. Due to the limit of detection and sensitivity of proteomics technologies, our simulated sets of true negatives may include false negatives (ORFs that actually are translated but not detected in proteomics, and thus may not reflect the actual biology of the organism). This is a limitation of our study, but improvements can be made in the future by searching our set of true negatives against other publically available high-coverage proteomics datasets, especially from PeptideAtlas and PRIDE.

Perhaps one advantage in translating all of our potential ORFs from all 6 frames of translation, given accurate DNA sequences for each chromosome, is

that all of the possible peptide sequences are considered, except for potential ORFs produced by splicing. The disadvantage is a large database size, but yeast has a relatively small genome size, this may not pose an issue. However, for organisms that have small draft genomes, such as newly discovered bacterial species, our method may be feasible and applicable. Another advantage is that every possible peptide sequence is considered since the method is not dependent on databases, such as RefSeq (Tatusova et al., 2014), that may not cover all sequences.

One disadvantage of using yeast is that it has very few splicing events, so it is not a comparable representation of human gene expression. Therefore, this is another limit of our method and study. But for other organisms where splicing occurs frequently, perhaps it would be worth running a splice site predictor to help inform where potential new exon-exon sequences might be found.

5.2.6 Jackknife Testing of the UAR and Proteomics Pipeline

One of the most limiting factors is the quality of the proteomics data, evidenced by the large increase in the average sensitivity (across all biological replicates in all 20 groups) in detection of protein-coding SGD genes to 97.0% after eliminating 2,987 SGD genes that were not detected in the original proteomics study by Tyagi and Pedrioli (2015). After this correction, the average sensitivity and specificity for detecting SGD proteins that were previously reported to be detectable were 97.0% and 99.8%, demonstrating that the pipeline

produced in this study yields results at a comparable level of quality to the original proteomics study.

Although the pipeline performed well in detection of known, previously detected, protein-coding SGD genes, only 110 of the 6,600 masked SGD genes were 'discovered' by the pipeline across all biological replicates for all 20 groups, yielding an average sensitivity rate of 1.6%. This discrepancy in sensitivity may be due to UAR ORF sequences not being accurate representations of actual peptides and proteins coded for. Therefore, searching against a database consisting of unrealistic sequences may not be as fruitful. Using gene model prediction to determine which UAR ORFs (or multiple adjacent UAR ORFs, for instance) are most likely to produce peptides and proteins may have produced a higher sensitivity in this case.

However, despite the areas of major improvement that exist within this study, the pipeline nonetheless does 'discover' 110 of the 6,600 masked SGD genes, which serves the objective of 'discovering' new peptides and proteins. In addition to the aforementioned refinements, UAR ORFs can be further filtered by using RNA-seq reads to help increase the number of discovered peptides and proteins. Even though the sensitivity is very low, the specificity is quite high at 98.6%, signifying that the sequences that are detected have a very high probability of being a real peptide or protein. For experiments in which suspected potential peptides or proteins are required to be detected in vivo, the high specificity will be advantageous in providing a small, high-quality set of sequences that require further testing.

5.2.7 Application of Proteogenomics Methods on Genome Annotation of Less Well-Annotated Species

In less well-annotated species with sequenced chromosomes, it would be feasible to 6-frame translate all chromosomes to find all possible open reading frames and therefore potential peptides and proteins and search these sequences against shotgun proteomics data. However, it might be difficult to delineate between false positive and true positive proteomics search results. In this study, we created two different sequence databases of *Saccharomyces* Genome Database known protein-coding genes: the Top 627 genes with the highest number of RNA-seq reads and the Representative sequences between 964 and 1178 bp, based on the median length of protein-coding genes. These databases contain a curated set of true positives - sequences that have ample evidence of being translated into proteins and are either well expressed on the RNA-seq level or are a representative sample of the lengths of the vast majority of yeast protein-coding genes. The search results of Top627 and Representative964_1178bp databases can be compared against those of the databases comprised of un-annotated region open reading frames (potential protein-coding sequences). The comparison can help inform search results against novel or potentially expressed sequences in less well-annotated species, to determine whether detected potential protein-coding sequences have characteristics resembling true positives or not, for example.

Within the main context of RNA-sequencing, the pipeline constructed in this study, along with previously mentioned refinements, may be applied to poorly

annotated genomes of less studied or newly discovered organisms in conjunction with software and methods from current literature. For example, DNA sequencing and RNA sequencing may be performed in a new organism. The transcriptome may then be assembled de novo from RNA-sequencing results by software, such as Cufflinks (Trapnell et al., 2010) or Trinity (Haas et al., 2013). Six-frame translation can be performed on results from both the DNA sequencing and de novo transcriptome assembly and compared to create a comprehensive set of un-annotated region open-reading frames (UAR ORFs). The entire set of UAR ORFs can be filtered by the results of gene model predictors, with the remaining sequences contained within a FASTA database. Proteomics can also be performed on the new organism, and the raw proteomics results can be searched against the aforementioned FASTA database. Proteomics searches may be refined by strategies recommended in Blakeley et al. (2012). In summary, the methods and pipeline created in this study should be used alongside existing bioinformatics tools to capitalise on their strengths and reduce the impact of their limitations.

5.3 Future Work

5.3.1 UAR-Pipeline

Currently, the UAR-Pipeline can process unstranded RNA-seq reads. However, to make the program more general and comprehensive, further work could be done to modify or add a module to the program to analyse strand-specific

reads as well. This new capability would allow the use of stranded data, which could help determine where RNA-seq reads from the unstranded dataset were transcribed from.

5.3.2 IGB Quickload Sites

Creating the Integrated Genome Browser Quickload Sites to include RNA-seq data, genome annotations, and conservation information can be applied to species in addition to *Saccharomyces cerevisiae*. The method of separating genome annotations into primary and secondary annotations may also be applicable to other species.

5.3.3 Proteomics Datasets

The set of UAR ORFs could have been searched against more than just one proteomics dataset. There are publically available repositories for raw proteomics datasets, such as PRIDE (Vizcaino et al., 2013) and PeptideAtlas (Desiere et al., 2006). Although many of the datasets may not have lower percentages of coverage, other peptides and proteins may have been detected that were not detected in the one dataset used in this study.

5.3.4 Proteomics Analysis

Tandem mass spectra searches against the peptide/protein sequence database led to peptides that matching multiple spectra and multiple proteins that were grouped together that were indistinguishable.

For better and more efficient proteomics searches, spectral processing before performing the sequence database spectral library searching may yield more accurate results. For example, low quality spectra may be discarded (Nesvizhskii et al., 2006; Gentzel et al., 2003). Moreover, redundant spectra can be clustered for higher efficiency (Beer et al., 2004; Tabb et al., 2005). Determination of charge state (Na et al., 2008; Sadygov et al., 2008) and peptide mass (Mayampurath et al., 2008; Shinkawa et al., 2009) may also be improved. In addition, de novo sequencing for unmatched MS/MS spectra can be performed (Nesvizhskii et al., 2006). Unrestrictive (blind) searches is another possible method in which all possible chemical or post-translational modifications are allowed or performing error-tolerant searches where mismatches between the peptide sequence producing the spectrum and the database sequence the spectrum was matched to (Dasari et al., 2010).

5.3.5 Experimental Validation of UAR ORF Targets

Both UAR ORF targets 17,069 and 24,011 have RNA-seq reads aligned to their genomic regions in the Stringent Alignment, strongly indicating that transcription is active at those locations. Proteomics analysis provided further evidence of expression through matching the targets' respective corresponding peptides with MS/MS spectra. These two critical pieces of evidence strongly suggest that there may be transcriptional activity at the genomic locations and translational activity for the corresponding mRNA molecules. Therefore, experimental validation in the laboratory would be an appropriate and logical

next step in order to determine whether these genomic regions are actually being expressed. If these regions are expressed, experimental analysis may also provide further information regarding what kind of peptides or small proteins are synthesised.

5.4 Summary of Conclusions

The main objective in this study was to use RNA-sequencing and proteomics to discover new genomic features in un-annotated regions in *Saccharomyces cerevisiae*. In this study, RNA-sequencing data were aligned at three different stringency levels to elucidate the complexities within the dataset, accounting for duplicate read alignments and those that mapped to multiple genomic locations. Moreover, *Saccharomyces* Genome Database annotations were divided into Primary and Secondary Annotations to use the former in defining the locations of un-annotated regions. These un-annotated regions were translated in six frames, the sequences of which were converted into a FASTA database of hypothetical peptide sequences. Proteomics data from Tyagi and Pedrioli (2015) were then searched against the FASTA database to detect hypothetical peptides. In a jackknife fashion, 6,600 SGD protein-coding genes were divided into 20 groups, each comprised of 6,270 unmasked genes with 330 masked genes to be discovered. Out of the 6,600 genes, 3,613 were detected in the original Tyagi and Pedrioli (2015) study. Across all groups, the average sensitivity and specificity for detecting the 3,613 unmasked SGD proteins genes were 97.0% and 99.8%, respectively. The average sensitivity and specificity for discovering

masked SGD genes across all 20 groups were 1.6% and 98.6%, respectively.

Bibliography

Aebersold, R., and D. R. Goodlett. 2001. Mass spectrometry in proteomics. *Chemical Reviews* 101(2):269–295.

Aebersold, Ruedi, and Matthias Mann. 2003. Mass spectrometry-based proteomics. *Nature* 422(6928):198–207.

Albert, Istvan, Travis N. Mavrich, Lynn P. Tomsho, Ji Qi, Sara J. Zanton, Stephan C. Schuster, and B. Franklin Pugh. 2007. Translational and rotational settings of H2a.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* 446(7135):572–576.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215(3):403–410.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17):3389–3402.

Anders, Simon, and Wolfgang Huber. 2010. Differential expression analysis for sequence count data. *Genome Biology* 11(10):R106.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2014. HTSeq - A Python framework to work with high-throughput sequencing data. *bioRxiv* 002824.

Andersen, Sabrina L., Roketa S. Sloan, Thomas D. Petes, and Sue Jinks-Robertson. 2015. Genome-destabilizing effects associated with top1 loss or accumulation of top1 cleavage complexes in yeast. *PLoS genetics* 11(4): e1005098.

Andreeva, Antonina, Dave Howorth, John-Marc Chandonia, Steven E. Brenner, Tim J. P. Hubbard, Cyrus Chothia, and Alexey G. Murzin. 2007. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research*.

Andrews, Shea J., and Joseph A. Rothnagel. 2014. Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics* 15(3):193–204.

Ansong, Charles, Samuel O. Purvine, Joshua N. Adkins, Mary S. Lipton, and Richard D. Smith. 2008. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Briefings in Functional Genomics & Proteomics* 7(1):50–62.

Apweiler, R., T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. R. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. A. Sigrist, and E.M. Zdobnov. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research* 29(1):37–40.

Arthur Lesk. 2008. *Introduction to Bioinformatics*. 3rd ed. Oxford University Press.

Attwood, Teresa K., Alain Coletta, Gareth Muirhead, Athanasia Pavlopoulou, Peter B. Philippou, Ivan Popov, Carlos Rom-Mateo, Athina Theodosiou, and Alex L. Mitchell. 2012. The PRINTS database: a fine-grained protein sequence annotation and analysis resourceits status in 2012. *Database* 2012: bas019.

Balakrishnan, Rama, Julie Park, Kalpana Karra, Benjamin C. Hitz, Gail Binkley, Eurie L. Hong, Julie Sullivan, Gos Micklem, and J. Michael Cherry. 2012. YeastMine–an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database: The Journal of Biological Databases and Curation* 2012:bar062.

Bateman, Alex, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam GriffithsJones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L. L. Sonnhammer, David J. Studholme, Corin Yeats, and Sean R. Eddy. 2004. The Pfam protein families database. *Nucleic Acids Research* 32(suppl 1):D138–D141.

Beer, Ilan, Eilon Barnea, Tamar Ziv, and Arie Admon. 2004. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* 4(4): 950–960.

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2013. GenBank. *Nucleic Acids Research* 41(Database issue):D36–42.

Bernstein, F. C., T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology* 112(3):535–542.

Blakeley, Paul, Ian M. Overton, and Simon J. Hubbard. 2012. Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. *Journal of Proteome Research* 11(11):5221–5234.

Blanchette, Mathieu, W. James Kent, Cathy Riemer, Laura Elnitski, Arian F. A. Smit, Krishna M. Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D. Green, David Haussler, and Webb Miller. 2004. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Research* 14(4):708–715.

Boeckmann, Brigitte, Marie-Claude Blatter, Livia Famiglietti, Ursula Hinz, Lydie Lane, Bernd Roechert, and Amos Bairoch. 2005. Protein variety and functional diversity: Swiss-Prot annotation in its biological context. *Comptes Rendus Biologies* 328(1011):882–899.

Boutet, Emmanuel, Damien Lieberherr, Michael Tognolli, Michel Schneider, Parit Bansal, AlanJ. Bridge, Sylvain Poux, Lydie Bougueleret, and Ioannis Xenarios. 2016. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. In *Plant Bioinformatics*, ed. David Edwards, 23–54. Methods in Molecular Biology 1374, Springer New York.

Bru, Catherine, Emmanuel Courcelle, Sbastien Carrre, Yoann Beausse, Sandrine Dalmar, and Daniel Kahn. 2005. The ProDom database of protein domain families: more emphasis on 3d. *Nucleic Acids Research* 33(suppl 1): D212–D215.

Buhler, Cyril, Valrie Borde, and Michael Lichten. 2007. Mapping Meiotic Single-Strand DNA Reveals a New Landscape of DNA Double-Strand Breaks in Saccharomyces cerevisiae. *PLoS Biology* 5(12).

Bullard, James H., Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94.

Burge, Chris, and Samuel Karlin. 1997. Prediction of complete gene structures in human genomic DNA1. *Journal of Molecular Biology* 268(1):78–94.

Campagna, Davide, Alessandro Albiero, Alessandra Bilardi, Elisa Caniato, Claudio Forcato, Svetlin Manavski, Nicola Vitulo, and Giorgio Valle. 2009. PASS: a program to align short sequences. *Bioinformatics (Oxford, England)* 25(7):967–968.

Cherry, J. Michael, Eurie L. Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T. Chan, Karen R. Christie, Maria C. Costanzo, Selina S. Dwight, Stacia R. Engel, Dianna G. Fisk, Jodi E. Hirschman, Benjamin C. Hitz, Kalpana Karra, Cynthia J. Krieger, Stuart R. Miyasato, Rob S. Nash, Julie Park, Marek S. Skrzypek, Matt Simison, Shuai Weng, and Edith D. Wong. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research* 40(D1):D700–D705.

Choi, Hyungwon, and Alexey I. Nesvizhskii. 2008. False Discovery Rates and Related Statistical Concepts in Mass Spectrometry-Based Proteomics. *Journal of Proteome Research* 7(1):47–50.

Clark, Tyson A., Charles W. Sugnet, and Manuel Ares. 2002. Genomewide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays. *Science* 296(5569):907–910.

Cloonan, Nicole, Alistair R. R. Forrest, Gabriel Kolle, Brooke B. A. Gardiner, Geoffrey J. Faulkner, Mellissa K. Brown, Darrin F. Taylor, Anita L. Steptoe, Shivangi Wani, Graeme Bethel, Alan J. Robertson, Andrew C. Perkins, Stephen J. Bruce, Clarence C. Lee, Swati S. Ranade, Heather E. Peckham, Jonathan M. Manning, Kevin J. McKernan, and Sean M. Grimmond. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* 5(7):613–619.

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Micha Wojciech Szczeniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biology* 17:13.

Conrads, Thomas P., Haleem J. Issaq, and Timothy D. Veenstra. 2002. New Tools for Quantitative Phosphoproteome Analysis. *Biochemical and Biophysical Research Communications* 290(3):885–890.

Consortium, The UniProt. 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research* 40(D1):D71–D75.

Darling, Aaron C. E., Bob Mau, Frederick R. Blattner, and Nicole T. Perna. 2004. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research* 14(7):1394–1403.

Darling, Aaron E., Bob Mau, and Nicole T. Perna. 2010. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE* 5(6):e11147.

Dasari, Surendra, Matthew C. Chambers, Robbert J. Slebos, Lisa J. Zimmerman, Amy-Joan L. Ham, and David L. Tabb. 2010. TagRecon: High-Throughput Mutation Identification through Sequence Tagging. *Journal of Proteome Research* 9(4):1716–1726.

David, Lior, Wolfgang Huber, Marina Granovskaia, Joern Toedling, Curtis J. Palm, Lee Bofkin, Ted Jones, Ronald W. Davis, and Lars M. Steinmetz. 2006. A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences* 103(14):5320–5325.

Delcher, Arthur L., Simon Kasif, Robert D. Fleischmann, Jeremy Peterson, Owen White, and Steven L. Salzberg. 1999. Alignment of whole genomes. *Nucleic Acids Research* 27(11):2369–2376.

Delcher, Arthur L., Adam Phillippy, Jane Carlton, and Steven L. Salzberg. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research* 30(11):2478–2483.

Desiere, Frank, Eric W. Deutsch, Nichole L. King, Alexey I. Nesvizhskii, Parag Mallick, Jimmy Eng, Sharon Chen, James Eddes, Sandra N. Loevenich, and Ruedi Aebersold. 2006. The PeptideAtlas project. *Nucleic Acids Research* 34(suppl 1):D655–D658.

Deutsch, Eric W., Luis Mendoza, David Shteynberg, Terry Farrah, Henry Lam, Natalie Tasman, Zhi Sun, Erik Nilsson, Brian Pratt, Bryan Prazen, Jimmy K. Eng, Daniel B. Martin, Alexey I. Nesvizhskii, and Ruedi Aebersold. 2010. A guided tour of the Trans-Proteomic Pipeline. *PROTEOMICS* 10(6):1150–1159.

van Dijk, E. L., C. L. Chen, Y. dAubenton Carafa, S. Gourvennec, M. Kwapisz, V. Roche, C. Bertrand, M. Silvain, P. Legoix-N, S. Loeillet, A. Nicolas, C. Thermes, and A. Morillon. 2011. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature* 475(7354):114–117.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1): 15–21.

Doerge, Rebecca W. 2002. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* 3(1):43–52.

Eaton, Matthew L., Kyriaki Galani, Sukhyun Kang, Stephen P. Bell, and David M. MacAlpine. 2010. Conserved nucleosome positioning defines replication origins. *Genes & Development* 24(8):748–753.

Elias, Joshua E., and Steven P. Gygi. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* 4(3):207–214.

Eng, Jimmy K., Bernd Fischer, Jonas Grossmann, and Michael J. MacCoss. 2008. A Fast SEQUEST Cross Correlation Algorithm. *Journal of Proteome Research* 7(10):4598–4602.

Eng, Jimmy K., Tahmina A. Jahan, and Michael R. Hoopmann. 2013. Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS* 13(1):22–24.

Eng, Jimmy K., Ashley L. McCormack, and John R. Yates III. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* 5(11):976–989.

Engstrm, Pr G., Tamara Steijger, Botond Sipos, Gregory R. Grant, Andr Kahles, The RGASP Consortium, Gunnar Rtsch, Nick Goldman, Tim J. Hubbard, Jennifer Harrow, Roderic Guig, and Paul Bertone. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods* 10(12):1185–1191.

Feng, Jian, Daniel Q. Naiman, and Bret Cooper. 2007. Probability-based pattern recognition and statistical framework for randomization: modeling tandem mass spectrum/peptide sequence false match frequencies. *Bioinformatics* 23(17):2210–2217.

Field, Yair, Noam Kaplan, Yvonne Fondufe-Mittendorf, Irene K. Moore, Eilon Sharon, Yaniv Lubling, Jonathan Widom, and Eran Segal. 2008. Distinct Modes of Regulation by Chromatin Encoded through Nucleosome Positioning Signals. *PLoS Comput Biol* 4(11):e1000216.

Finn, Robert D., Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L. L. Sonnhammer, John Tate, and Marco Punta. 2014. Pfam: the protein families database. *Nucleic Acids Research* 42(D1):D222–D230.

Flicek, Paul, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos Garca Girn, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Nathan Johnson, Thomas Juettemann, Andreas K. Khri, Stephen Keenan, Eugene Kulesha, Fergal J. Martin, Thomas Maurel, William M. McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet S. Riat, Magali Ruffier, Daniel Sheppard, Kieron Taylor, Anja Thormann, Stephen J. Trevanion, Alessandro Vullo, Steven P. Wilder, Mark Wilson, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Jennifer Harrow, Javier Herrero, Tim J. P. Hubbard, Rhoda Kinsella, Matthieu Muffato, Anne Parker, Giulietta Spudich, Andy Yates, Daniel R. Zerbino, and Stephen M. J. Searle. 2014. Ensembl 2014. *Nucleic Acids Research* 42(D1):D749–D755.

Flicek, Paul, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Leo Gordon, Maurice Hendrix, Thibaut Hourlier, Nathan Johnson, Andreas Khri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Eugene Kulesha, Pontus Larsson, Ian Longden, William McLaren, Bert Overduin, Bethan Pritchard, Harpreet Singh Riat, Daniel Rios, Graham R. S. Ritchie, Magali Ruffier, Michael Schuster, Daniel Sobral, Giulietta Spudich, Y. Amy Tang, Stephen Trevanion, Jana Vandrovcova, Albert J. Vilella, Simon White, Steven P. Wilder, Amonida Zadissa, Jorge Zamora, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xos M. Fernndez-Suarez, Javier Herrero, Tim J. P. Hubbard, Anne Parker, Glenn Proctor, Jan Vogel, and Stephen M. J. Searle. 2011. Ensembl 2011. *Nucleic Acids Research* 39(Database issue):D800–806.

Gentzel, Marc, Thomas Kcher, Saravanan Ponnusamy, and Matthias Wilm. 2003. Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics* 3(8):1597–1610.

Ghaemmaghami, Sina, Won-Ki Huh, Kiowa Bower, Russell W. Howson, Archana Belle, Noah Dephoure, Erin K. O'Shea, and Jonathan S. Weissman. 2003. Global analysis of protein expression in yeast. *Nature* 425(6959): 737–741.

Gingeras Group. 2016. STAR: Spliced Transcripts Alignments and Reconstruction - Introduction.

Goh, Phuay-Yee, and Uttam Surana. 1999. Cdc4, a Protein Required for the Onset of S Phase, Serves an Essential Function during G2/M Transition in Saccharomyces cerevisiae. *Molecular and Cellular Biology* 19(8):5512–5522.

Goujon, Mickael, Hamish McWilliam, Weizhong Li, Franck Valentin, Silvano Squizzato, Juri Paern, and Rodrigo Lopez. 2010. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Research* 38(Web Server issue):W695–699.

Granovskaia, Marina V., Lars J. Jensen, Matthew E. Ritchie, Joern Toedling, Ye Ning, Peer Bork, Wolfgang Huber, and Lars M. Steinmetz. 2010. High-resolution transcription atlas of the mitotic cell cycle in budding yeast. *Genome Biology* 11(3):R24.

Gygi, Steven P., Beate Rist, Scott A. Gerber, Frantisek Turecek, Michael H. Gelb, and Ruedi Aebersold. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology* 17(10): 994–999.

Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D. MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N. Dewey, Robert Henschel, Richard D. LeDuc, Nir Friedman, and Aviv Regev. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8(8):1494–1512.

Haft, Daniel H., Jeremy D. Selengut, Roland A. Richter, Derek Harkins, Malay K. Basu, and Erin Beck. 2013. TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Research* 41(D1):D387–D395.

Hager, James W. 2002. A new linear ion trap mass spectrometer. *Rapid Communications in Mass Spectrometry* 16(6):512–526.

Hardcastle, Thomas J., and Krystyna A. Kelly. 2010. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 11:422.

Hesselberth, Jay R., Xiaoyu Chen, Zhihong Zhang, Peter J. Sabo, Richard Sandstrom, Alex P. Reynolds, Robert E. Thurman, Shane Neph, Michael S.

Kuehn, William S. Noble, Stanley Fields, and John A. Stamatoyannopoulos. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods* 6(4):283–289.

Hontz, Robert D., Rachel O. Niederer, Joseph M. Johnson, and Jeffrey S. Smith. 2009. Genetic identification of factors that modulate ribosomal DNA transcription in Saccharomyces cerevisiae. *Genetics* 182(1):105–119.

Houseley, Jonathan, and David Tollervey. 2010. Apparent Non-Canonical Trans-Splicing Is Generated by Reverse Transcriptase In Vitro. *PLoS ONE* 5(8).

Hunter, J.D. 2007. Matplotlib: A 2d Graphics Environment. *Computing in Science Engineering* 9(3):90–95.

Illumina. 2011. Paired-End Sample Preparation Guide.

———. 2013. TruSeq Stranded mRNA Sample Preparation Guide.

———. 2014. TruSeq RNA Sample Preparation v2 Guide.

———. 2016. An Introduction to Next-Generation Sequencing Technology.

Ingolia, Nicholas T., Sina Ghaemmaghami, John R. S. Newman, and Jonathan S. Weissman. 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* 324(5924): 218–223.

Jean, Graldine, Andr Kahles, Vipin T. Sreedharan, Fabio De Bona, and Gunnar Rtsch. 2010. RNA-Seq read alignments with PALMapper. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* Chapter 11:Unit 11.6.

Jiang, Lichun, Felix Schlesinger, Carrie A. Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R. Gingeras, and Brian Oliver. 2011. Synthetic spike-in standards for RNA-seq experiments. *Genome Research* 21(9):1543–1551.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236–1240.

Kaisers, Wolfgang, Holger Schwender, and Heiner Schaal. 2014. Hierarchical clustering of DNA k-mer counts in RNA-seq fastq files reveals batch effects. *arXiv:1405.0114 [q-bio]*. ArXiv: 1405.0114.

Kapranov, Philipp, Jill Cheng, Sujit Dike, David A. Nix, Radharani Duttagupta, Aarron T. Willingham, Peter F. Stadler, Jana Hertel, Jrg Hackermller, Ivo L. Hofacker, Ian Bell, Evelyn Cheung, Jorg Drenkow, Erica

Dumais, Sandeep Patel, Gregg Helt, Madhavan Ganesh, Srinka Ghosh, Antonio Piccolboni, Victor Sementchenko, Hari Tammana, and Thomas R. Gingeras. 2007. RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science* 316(5830):1484–1488.

Keller, Andrew, Alexey I. Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. 2002. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Analytical Chemistry* 74(20):5383–5392.

Kent, W. James, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and and David Haussler. 2002. The Human Genome Browser at UCSC. *Genome Research* 12(6):996–1006.

Kertesz, Michael, Yue Wan, Elad Mazor, John L. Rinn, Robert C. Nutter, Howard Y. Chang, and Eran Segal. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467(7311):103–107.

Kessler, Marco M., Qiandong Zeng, Sarah Hogan, Robin Cook, Arturo J. Morales, and Guillaume Cottarel. 2003. Systematic Discovery of New Genes in the Saccharomyces cerevisiae Genome. *Genome Research* 13(2):264–271.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* 14(4):R36.

King, Nichole L., Eric W. Deutsch, Jeffrey A. Ranish, Alexey I. Nesvizhskii, James S. Eddes, Parag Mallick, Jimmy Eng, Frank Desiere, Mark Flory, Daniel B. Martin, Bong Kim, Hookeun Lee, Brian Raught, and Ruedi Aebersold. 2006. Analysis of the Saccharomyces cerevisiae proteome with PeptideAtlas. *Genome Biology* 7(11):R106.

Kosuge, Takehide, Jun Mashima, Yuichi Kodama, Takatomo Fujisawa, Eli Kaminuma, Osamu Ogasawara, Kousaku Okubo, Toshihisa Takagi, and Yasukazu Nakamura. 2014. DDBJ progress report: a new submission system for leading to a correct annotation. *Nucleic Acids Research* 42(D1):D44–D49.

Krogh, Anders, Michael Brown, I. Saira Mian, Kimmen Sjlander, and David Haussler. 1994. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *Journal of Molecular Biology* 235(5):1501–1531.

Kuai, Letian, Feng Fang, J. Scott Butler, and Fred Sherman. 2004. Polyadenylation of rRNA in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* 101(23):8581–8586.

Kumar, Navin, Ramesh C. Meena, and Amitabha Chakrabarti. 2011. Over-Expression of YLR162w in Saccharomyces cerevisiae Inhibits Cell Proliferation and Renders Cells Susceptible to the Hypoxic Conditions Induced by Cobalt Chloride. *Indian Journal of Microbiology* 51(2):206–211.

Kurtz, Stefan, Adam Phillippy, Arthur L. Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L. Salzberg. 2004. Versatile and open software for comparing large genomes. *Genome Biology* 5(2):R12.

Langmead, Ben, and Steven L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9(4):357–359.

Lardenois, Aurlie, Yuchen Liu, Thomas Walther, Frdric Chalmel, Bertrand Evrard, Marina Granovskaia, Angela Chu, Ronald W. Davis, Lars M. Steinmetz, and Michael Primig. 2011. Execution of the meiotic noncoding RNA expression program and the onset of gametogenesis in yeast require the conserved exosome subunit Rrp6. *Proceedings of the National Academy of Sciences* 108(3):1058–1063.

Lee, Rosalind C., Rhonda L. Feinbaum, and Victor Ambros. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75(5):843–854.

Lee, William, Desiree Tillo, Nicolas Bray, Randall H. Morse, Ronald W. Davis, Timothy R. Hughes, and Corey Nislow. 2007. A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genetics* 39(10):1235–1244.

Lees, Jonathan, Corin Yeats, James Perkins, Ian Sillitoe, Robert Rentzsch, Benoit H. Dessailly, and Christine Orengo. 2012. Gene3d: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Research* 40(D1):D465–D471.

Letunic, Ivica, Tobias Doerks, and Peer Bork. 2015. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Research* 43(D1):D257–D260.

Li, Heng, and Richard Durbin. 2009. Fast and accurate short read alignment with BurrowsWheeler transform. *Bioinformatics* 25(14):1754–1760.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Li, Heng, and Nils Homer. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics* 11(5):473–483.

Lipton, Mary S., Ljiljana Pasa-Tolic', Gordon A. Anderson, David J. Anderson, Deanna L. Auberry, John R. Battista, Michael J. Daly, Jim Fredrickson, Kim K. Hixson, Heather Kostandarithes, Christophe Masselon, Lye Meng Markillie, Ronald J. Moore, Margaret F. Romine, Yufeng Shen, Eric Stritmatter, Nikola Tolic', Harold R. Udseth, Amudhan Venkateswaran, Kwong-Kwok Wong, Rui Zhao, and Richard D. Smith. 2002. Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. *Proceedings of the National Academy of Sciences of the United States of America* 99(17):11049–11054.

Lister, Ryan, Ronan C. O'Malley, Julian Tonti-Filippini, Brian D. Gregory, Charles C. Berry, A. Harvey Millar, and Joseph R. Ecker. 2008. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* 133(3):523–536.

Lovn, Jakob, David A. Orlando, Alla A. Sigova, Charles Y. Lin, Peter B. Rahl, Christopher B. Burge, David L. Levens, Tong Ihn Lee, and Richard A. Young. 2012. Revisiting global gene expression analysis. *Cell* 151(3):476–482.

Lunter, Gerton, and Martin Goodson. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* 21(6):936–939.

MacIsaac, Kenzie D., Ting Wang, D. Benjamin Gordon, David K. Gifford, Gary D. Stormo, and Ernest Fraenkel. 2006. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* 7(1): 113.

Madoui, Mohammed-Amin, Stefan Engelen, Corinne Cruaud, Caroline Belser, Laurie Bertrand, Adriana Alberti, Arnaud Lemainque, Patrick Wincker, and Jean-Marc Aury. 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* 16:327.

Mancera, Eugenio, Richard Bourgon, Alessandro Brozzi, Wolfgang Huber, and Lars M. Steinmetz. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454(7203):479–485.

Mann, M., and M. Wilm. 1994. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Analytical Chemistry* 66(24): 4390–4399.

Marco-Sola, Santiago, Michael Sammeth, Roderic Guig, and Paolo Ribeca. 2012. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods* 9(12):1185–1188.

Margaret O. Dayhoff, ed. 1969. *Atlas of Protein Sequence and Structure*. Silver Spring, Maryland: National Biomedical Research Foundation.

Marshall, A. G., C. L. Hendrickson, and G. S. Jackson. 1998. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrometry Reviews* 17(1):1–35.

Martin, S. E., J. Shabanowitz, D. F. Hunt, and J. A. Marto. 2000. Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry. *Analytical Chemistry* 72(18):4266–4274.

MasonLab. 2016. TBAG - Icbwiki.

Mavrich, Travis N., Ilya P. Ioshikhes, Bryan J. Venters, Cizhong Jiang, Lynn P. Tomsho, Ji Qi, Stephan C. Schuster, Istvan Albert, and B. Franklin Pugh. 2008. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research*.

Mayampurath, Anoop M., Navdeep Jaitly, Samuel O. Purvine, Matthew E. Monroe, Kenneth J. Auberry, Joshua N. Adkins, and Richard D. Smith. 2008. DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics (Oxford, England)* 24(7):1021–1023.

Mayer, Andreas, Michael Lidschreiber, Matthias Siebert, Kristin Leike, Johannes Sding, and Patrick Cramer. 2010. Uniform transitions of the general RNA polymerase II transcription complex. *Nature Structural & Molecular Biology* 17(10):1272–1278.

Mirgorodskaya, O. A., Y. P. Kozmin, M. I. Titov, R. Krner, C. P. Snksen, and P. Roepstorff. 2000. Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards. *Rapid communications in mass spectrometry: RCM* 14(14):1226–1232.

Miura, Fumihito, Noriko Kawaguchi, Jun Sese, Atsushi Toyoda, Masahira Hattori, Shinichi Morishita, and Takashi Ito. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proceedings of the National Academy of Sciences* 103(47):17846–17851.

Morin, Ryan. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* 45(1):81–94.

Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5(7):621–628.

Mortimer, Robert K., and John R. Johnston. 1986. Genealogy of Principal Strains of the Yeast Genetic Stock Center. *Genetics* 113(1):35–43.

Mudge, Jonathan M., Adam Frankish, and Jennifer Harrow. 2013. Functional transcriptomics in the post-ENCODE era. *Genome Research*.

Mullins, Michael, Laurent Perreard, John F. Quackenbush, Nicholas Gauthier, Steven Bayer, Matthew Ellis, Joel Parker, Charles M. Perou, Aniko Szabo, and Philip S. Bernard. 2007. Agreement in Breast Cancer Classification between Microarray and Quantitative Reverse Transcription PCR from Fresh-Frozen and Formalin-Fixed, Paraffin-Embedded Tissues. *Clinical Chemistry* 53(7):1273–1279.

Na, Seungjin, Eunok Paek, and Cheolju Lee. 2008. CIFTER: Automated Charge-State Determination for Peptide Tandem Mass Spectra. *Analytical Chemistry* 80(5):1520–1528.

Nagalakshmi, Ugrappa, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* 320(5881): 1344–1349.

Neil, Helen, Christophe Malabat, Yves dAubenton Carafa, Zhenyu Xu, Lars M. Steinmetz, and Alain Jacquier. 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* 457(7232):1038–1042.

Nesvizhskii, Alexey I. 2010. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics* 73(11):2092–2123.

———. 2014. Proteogenomics: concepts, applications, and computational strategies. *Nature methods* 11(11):1114–1125.

Nesvizhskii, Alexey I., Andrew Keller, Eugene Kolker, and Ruedi Aebersold. 2003. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Analytical Chemistry* 75(17):4646–4658.

Nesvizhskii, Alexey I., Franz F. Roos, Jonas Grossmann, Mathijs Vogelzang, James S. Eddes, Wilhelm Gruissem, Sacha Baginsky, and Ruedi Aebersold. 2006. Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data Toward More Efficient Identification of Post-translational Modifications, Sequence Polymorphisms, and Novel Peptides. *Molecular & Cellular Proteomics* 5(4):652–670.

Nicol, John W., Gregg A. Helt, Steven G. Blanchard, Archana Raja, and Ann E. Loraine. 2009. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* 25(20):2730–2731.

Nikolskaya, Anastasia N., Cecilia N. Arighi, Hongzhan Huang, Winona C. Barker, and Cathy H. Wu. 2007. PIRSF Family Classification System for Protein Functional and Evolutionary Analysis. *Evolutionary Bioinformatics Online* 2:197–209.

Nookaew, Intawat, Marta Papini, Natapol Pornputtapong, Gionata Scalcinati, Linn Fagerberg, Matthias Uhln, and Jens Nielsen. 2012. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Research* 40(20):10084–10097.

Olarerin-George, Anthony O., and John B. Hogenesch. 2015. Assessing the prevalence of mycoplasma contamination in cell culture via a survey of NCBI's RNA-seq archive. *Nucleic Acids Research* 43(5):2535–2542.

Ong, Shao-En, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. 2002. Stable Isotope

Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Molecular & Cellular Proteomics* 1(5): 376–386.

Ozsolak, Fatih, Philipp Kapranov, Sylvain Foissac, Sang Woo Kim, Elane Fishilevich, A. Paula Monaghan, Bino John, and Patrice M. Milos. 2010. Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation. *Cell* 143(6):1018–1029.

Ozsolak, Fatih, Adam R. Platt, Dan R. Jones, Jeffrey G. Reifenberger, Lauryn E. Sass, Peter McInerney, John F. Thompson, Jayson Bowers, Mirna Jarosz, and Patrice M. Milos. 2009. Direct RNA sequencing. *Nature* 461(7265):814–818.

Pan, Jing, Mariko Sasaki, Ryan Kniewel, Hajime Murakami, Hannah G. Blitzblau, Sam E. Tischfield, Xuan Zhu, Matthew J. Neale, Maria Jasin, Nicholas D. Socci, Andreas Hochwagen, and Scott Keeney. 2011. A Hierarchical Combination of Factors Shapes the Genome-wide Topography of Yeast Meiotic Recombination Initiation. *Cell* 144(5):719–731.

Parenteau, Julie, Mathieu Durand, Steeve Vronneau, Andre-Anne Lacombe, Genevive Morin, Valrie Gurin, Bojana Cecez, Julien Gervais-Bird, Chu-Shin Koh, David Brunelle, Raymund J. Wellinger, Benoit Chabot, and Sherif Abou Elela. 2008. Deletion of Many Yeast Introns Reveals a Minority of Genes that Require Splicing for Function. *Molecular Biology of the Cell* 19(5):1932–1941.

Pedrioli, Patrick G. A. 2010. Trans-proteomic pipeline: a pipeline for proteomic analysis. *Methods in Molecular Biology (Clifton, N.J.)* 604:213–238.

Pedrioli, Patrick G. A., Jimmy K. Eng, Robert Hubley, Mathijs Vogelzang, Eric W. Deutsch, Brian Raught, Brian Pratt, Erik Nilsson, Ruth H. Angeletti, Rolf Apweiler, Kei Cheung, Catherine E. Costello, Henning Hermjakob, Sequin Huang, Randall K. Julian, Eugene Kapp, Mark E. McComb, Stephen G. Oliver, Gilbert Omenn, Norman W. Paton, Richard Simpson, Richard Smith, Chris F. Taylor, Weimin Zhu, and Ruedi Aebersold. 2004. A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology* 22(11):1459–1466.

Pedruzzi, Ivo, Catherine Rivoire, Andrea H. Auchincloss, Elisabeth Coudert, Guillaume Keller, Edouard de Castro, Delphine Baratin, Batrice A. Cuche, Lydie Bougueleret, Sylvain Poux, Nicole Redaschi, Ioannis Xenarios, Alan Bridge, and UniProt Consortium. 2013. HAMAP in 2013, new developments in the protein family classification and annotation system. *Nucleic Acids Research* 41(Database issue):D584–589.

Perkins, D. N., D. J. Pappin, D. M. Creasy, and J. S. Cottrell. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20(18):3551–3567.

Piehowski, Paul D., Vladislav A. Petyuk, Daniel J. Orton, Fang Xie, Manuel Ramirez-Restrepo, Anzhelika Engel, Andrew P. Lieberman, Roger L. Albin, David G. Camp, Richard D. Smith, and Amanda J. Myers. 2013. Sources of Technical Variability in Quantitative LC-MS Proteomics: Human Brain Tissue Sample Analysis. *Journal of proteome research* 12(5):2128–2137.

Qi, Ji, Asela J. Wijeratne, Lynn P. Tomsho, Yi Hu, Stephan C. Schuster, and Hong Ma. 2009. Characterization of meiotic crossovers and gene conversion by whole-genome sequencing in Saccharomyces cerevisiae. *BMC Genomics* 10(1):475.

Qing, Tao, Ying Yu, TingTing Du, and LeMing Shi. 2013. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. *Science China Life Sciences* 56(2):134–142.

Quail, Michael A., Miriam Smith, Paul Coupland, Thomas D. Otto, Simon R. Harris, Thomas R. Connor, Anna Bertoni, Harold P. Swerdlow, and Yong Gu. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13:341.

Rhee, Ho Sung, and B. Franklin Pugh. 2011. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* 147(6):1408–1419.

———. 2012. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* 483(7389):295–301.

Rhoads, Anthony, and Kin Fai Au. 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* 13(5):278–289.

Ribas, Juan Carlos, and Reed B. Wickner. 1998. The Gag Domain of the Gag-Pol Fusion Protein Directs Incorporation into the L-A Double-stranded RNA Viral Particles inSaccharomyces cerevisiae. *Journal of Biological Chemistry* 273(15):9306–9311.

Rice, Peter, Ian Longden, and Alan Bleasby. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16(6):276–277.

Rigden, Daniel J, Luciane V Mello, and Michael Y Galperin. 2004. The PA14 domain, a conserved all- domain in bacterial toxins, enzymes, adhesins and signaling molecules. *Trends in Biochemical Sciences* 29(7):335–339.

Risso, Davide, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. 2011. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics* 12: 480.

Rivera, Maria C., Bruce Maguire, and James A. Lake. 2015. Isolation of Ribosomes and Polysomes. *Cold Spring Harbor Protocols* 2015(3): pdb.prot081331.

Roberts, J D, B D Preston, L A Johnston, A Soni, L A Loeb, and T A Kunkel. 1989. Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro. *Molecular and Cellular Biology* 9(2): 469–476.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* 26(1):139–140.

Sadygov, Rovshan G., Zhiqi Hao, and Andreas F. R. Huhmer. 2008. Charger: combination of signal processing and statistical learning algorithms for precursor charge-state determination from electron-transfer dissociation spectra. *Analytical Chemistry* 80(2):376–386.

Sayols, Sergi, Denise Scherzinger, and Holger Klein. 2016. dupRadar: a Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. *BMC Bioinformatics* 17:428.

Schurch, Nicholas J., Pieta Schofield, Marek Gierliski, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G. Simpson, Tom Owen-Hughes, Mark Blaxter, and Geoffrey J. Barton. 2016. Evaluation of tools for differential gene expression analysis by RNA-seq on a 48 biological replicate experiment. *RNA* 22(6):839–851. ArXiv: 1505.02017.

Schwartz, Jae C, Michael W Senko, and John E. P Syka. 2002. A two-dimensional quadrupole ion trap mass spectrometer. *Journal of the American Society for Mass Spectrometry* 13(6):659–669.

Schwartz, Scott, W. James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C. Hardison, David Haussler, and Webb Miller. 2003. HumanMouse Alignments with BLASTZ. *Genome Research* 13(1):103–107.

Seqc/Maqc-Iii Consortium. 2014. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology* 32(9):903–914.

Shinkawa, Takashi, Kohji Nagano, Noriyuki Inomata, and Masayuki Haramura. 2009. A software program for more reliable precursor ion assignation from LC-MS analysis using LTQ ultra zoom scan. *Journal of Proteomics* 73(2):357–360.

Shteynberg, David, Eric W. Deutsch, Henry Lam, Jimmy K. Eng, Zhi Sun, Natalie Tasman, Luis Mendoza, Robert L. Moritz, Ruedi Aebersold, and Alexey I. Nesvizhskii. 2011. iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates. *Molecular & Cellular Proteomics* 10(12):M111.007690.

Siepel, Adam, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A.

Gibbs, W. James Kent, Webb Miller, and David Haussler. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* 15(8):1034–1050.

Sigrist, Christian J. A., Edouard de Castro, Lorenzo Cerutti, Batrice A. Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. 2013. New and continuing developments at PROSITE. *Nucleic Acids Research* 41(D1):D344–D347.

Sillitoe, Ian, Tony E. Lewis, Alison Cuff, Sayoni Das, Paul Ashford, Natalie L. Dawson, Nicholas Furnham, Roman A. Laskowski, David Lee, Jonathan G. Lees, Sonja Lehtinen, Romain A. Studer, Janet Thornton, and Christine A. Orengo. 2015. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research* 43(D1):D376–D381.

Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 147(1):195–197.

Stanke, Mario, and Stephan Waack. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19(suppl_2): ii215–ii225.

Stephens, Camilla, Stuart J. Harrison, Kemal Kazan, Frank W. N. Smith, Ken C. Goulter, Donald J. Maclean, and John M. Manners. 2005. Altered fungal sensitivity to a plant antimicrobial peptide through over-expression of yeast cDNAs. *Current Genetics* 47(3):194–201.

Stoesser, Guenter, Wendy Baker, Alexandra van den Broek, Evelyn Camon, Maria Garcia-Pastor, Carola Kanz, Tamara Kulikova, Rasko Leinonen, Quan Lin, Vincent Lombard, Rodrigo Lopez, Nicole Redaschi, Peter Stoehr, Mary Ann Tuli, Katerina Tzouvara, and Robert Vaughan. 2002. The EMBL Nucleotide Sequence Database. *Nucleic Acids Research* 30(1):21–26.

Tabb, David L., Melissa R. Thompson, Gurusahai Khalsa-Moyers, Nathan C. VerBerkmoes, and W. Hayes McDonald. 2005. MS2grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *Journal of the American Society for Mass Spectrometry* 16(8):1250–1261.

Tachibana, Christine, Jane Y. Yoo, Jean-Basco Tagne, Nataly Kacherovsky, Tong I. Lee, and Elton T. Young. 2005. Combined Global Localization Analysis and Transcriptome Data Identify Genes That Are Directly Coregulated by Adr1 and Cat8. *Molecular and Cellular Biology* 25(6):2138–2146.

Tarazona, Sonia, Fernando Garca-Alcalde, Joaqun Dopazo, Alberto Ferrer, and Ana Conesa. 2011. Differential expression in RNA-seq: a matter of depth. *Genome Research* 21(12):2213–2223.

Tatusov, Roman L., Eugene V. Koonin, and David J. Lipman. 1997. A Genomic Perspective on Protein Families. *Science* 278(5338):631–637.

Tatusova, Tatiana, Stacy Ciufo, Boris Fedorov, Kathleen O'Neill, and Igor Tolstoy. 2014. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Research* 42(Database issue):D553–559.

Thomas, Paul D., Michael J. Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. 2003. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Research* 13(9):2129–2141.

Tisseur, Mathieu, Marta Kwapisz, and Antonin Morillon. 2011. Pervasive transcription Lessons from yeast. *Biochimie* 93(11):1889–1896.

Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105–1111.

Trapnell, Cole, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* 28(5):511–515.

Tuller, Tamir, Eytan Ruppin, and Martin Kupiec. 2009. Properties of untranslated regions of the S. cerevisiae genome. *BMC Genomics* 10:391.

Tyagi, Kshitiz, and Patrick G. A. Pedrioli. 2015. Protein degradation and dynamic tRNA thiolation fine-tune translation at elevated temperatures. *Nucleic Acids Research* 43(9):4701–4712.

Valaskovic, G. A., N. L. Kelleher, and F. W. McLafferty. 1996. Attomole protein characterization by capillary electrophoresis-mass spectrometry. *Science (New York, N.Y.)* 273(5279):1199–1202.

Velculescu, Victor E., Lin Zhang, Wei Zhou, Jacob Vogelstein, Munira A. Basrai, Douglas E. Bassett, Phil Hieter, Bert Vogelstein, and Kenneth W. Kinzler. 1997. Characterization of the Yeast Transcriptome. *Cell* 88(2): 243–251.

Venters, Bryan J., Shinichiro Wachi, Travis N. Mavrich, Barbara E. Andersen, Peony Jena, Andrew J. Sinnamon, Priyanka Jain, Noah S. Rolleri, Cizhong Jiang, Christine Hemeryck-Walsh, and B. Franklin Pugh. 2011. A Comprehensive Genomic Binding Map of Gene and Chromatin Regulatory Proteins in Saccharomyces. *Molecular Cell* 41(4):480–492.

Vidgren, Virve, and John Londesborough. 2011. 125th Anniversary Review: Yeast Flocculation and Sedimentation in Brewing. *Journal of the Institute of Brewing* 117(4):475–487.

Vizcaino, J. A., R. G. Cote, A. Csordas, J. A. Dianes, A. Fabregat, J. M. Foster, J. Griss, E. Alpi, M. Birim, J. Contell, G. O'Kelly, A. Schoenegger,

D. Ovelleiro, Y. Perez-Riverol, F. Reisinger, D. Rios, R. Wang, and H. Hermjakob. 2013. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research* 41(D1):D1063–D1069.

Waern, Karl, and Michael Snyder. 2013. Extensive Transcript Diversity and Novel Upstream Open Reading Frame Regulation in Yeast. *G3: Genes|Genomes|Genetics* 3(2):343–352.

van der Walt, S., S.C. Colbert, and G. Varoquaux. 2011. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering* 13(2):22–30.

Wang, Kai, Darshan Singh, Zheng Zeng, Stephen J. Coleman, Yan Huang, Gleb L. Savich, Xiaping He, Piotr Mieczkowski, Sara A. Grimm, Charles M. Perou, James N. MacLeod, Derek Y. Chiang, Jan F. Prins, and Jinze Liu. 2010. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research* 38(18):e178.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10(1):57–63.

Wickham, Hadley. 2009. *ggplot2: elegant graphics for data analysis*. Springer New York.

Wightman, Bruce, Ilho Ha, and Gary Ruvkun. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* 75(5):855–862.

Wilson, Derek, Ralph Pethica, Yiduo Zhou, Charles Talbot, Christine Vogel, Martin Madera, Cyrus Chothia, and Julian Gough. 2009. SUPERFAMILYsophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research* 37(suppl 1):D380–D386.

Wu, Cathy H., Lai-Su L. Yeh, Hongzhan Huang, Leslie Arminski, Jorge Castro-Alvear, Yongxing Chen, Zhangzhi Hu, Panagiotis Kourtesis, Robert S. Ledley, Baris E. Suzek, C.R. Vinayaka, Jian Zhang, and Winona C. Barker. 2003. The Protein Information Resource. *Nucleic Acids Research* 31(1): 345–347.

Wu, Jia Qian, David Shteynberg, Manimozhiyan Arumugam, Richard A. Gibbs, and Michael R. Brent. 2004. Identification of Rat Genes by TWINSCAN Gene Prediction, RTPCR, and Direct Sequencing. *Genome Research* 14(4):665–671.

Wu, Thomas D., and Serban Nacu. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26(7):873–881.

Xu, Weihong, Jennifer G. Aparicio, Oscar M. Aparicio, and Simon Tavar. 2006. Genome-wide mapping of ORC and Mcm2p binding sites on tiling arrays and identification of essential ARS consensus sequences in S. cerevisiae. *BMC Genomics* 7(1):276.

Xu, Zhenyu, Wu Wei, Julien Gagneur, Fabiana Perocchi, Sandra Clauder-Mnster, Jurgi Camblong, Elisa Guffanti, Franoise Stutz, Wolfgang Huber, and Lars M. Steinmetz. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* 457(7232):1033–1037.

Yakovchuk, Peter, Ekaterina Protozanova, and Maxim D. Frank-Kamenetskii. 2006. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research* 34(2):564–574.

Yao, Xudong, Amy Freas, Javier Ramirez, Plamen A. Demirev, and Catherine Fenselau. 2001. Proteolytic 18o Labeling for Comparative Proteomics: Model Studies with Two Serotypes of Adenovirus. *Analytical Chemistry* 73(13):2836–2842.

Yassour, Moran, Tommy Kaplan, Hunter B. Fraser, Joshua Z. Levin, Jenna Pfiffner, Xian Adiconis, Gary Schroth, Shujun Luo, Irina Khrebtukova, Andreas Gnirke, Chad Nusbaum, Dawn-Anne Thompson, Nir Friedman, and Aviv Regev. 2009. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proceedings of the National Academy of Sciences* 106(9):3264–3269.

Yassour, Moran, Jenna Pfiffner, Joshua Z. Levin, Xian Adiconis, Andreas Gnirke, Chad Nusbaum, Dawn-Anne Thompson, Nir Friedman, and Aviv Regev. 2010. Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. *Genome Biology* 11(8):R87.

Zhang, Zhao, William E Theurkauf, Zhiping Weng, and Phillip D Zamore. 2012. Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly(A) selection. *Silence* 3:9.

Zhang, Zhihong, and Fred S. Dietrich. 2005. Mapping of transcription start sites in Saccharomyces cerevisiae using 5 SAGE. *Nucleic Acids Research* 33(9):2838–2851.

Zhou, Huilin, Jeffrey A. Ranish, Julian D. Watts, and Ruedi Aebersold. 2002. Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nature Biotechnology* 20(5):512–515.

Appendix A

Primary Annotations with their descriptions and publications from which they were produced. Descriptions of Annotations are directly from SGD (Cherry et al., 2012). Each track is in GFF3 format unless specified (e.g. bedgraph).

Track Name	**Description of Annotation**	**Publication**
David_2006_polyA_RNA_transcripts	segments resulting from the segmentation of the hybridization signal along genomic coordinates for poly(A) RNA	David et al. (2006)
David_2006_total_RNA_transcripts	segments resulting from the segmentation of the hybridization signal along genomic coordinates for total RNA	David et al. (2006)
Granovskaia_2010_antisense_transcripts	523 antisense transcripts identified in their study using A-AFFY-42 Affymetrix GeneChip S. cerevisiae tiling arrays	Granovskaia et al. (2010)
Granovskaia_2010_mitotic_ORFs	592 cyclic ORFs identified in the Granovskaia et al study using A-AFFY-42 Affymetrix GeneChip S. cerevisiae tiling arrays	Granovskaia et al. (2010)
Granovskaia_2010_novel_intergenic_transcripts	135 unannotated intergenic transcripts identified in the Granovskaia et al. study using A-AFFY-42 Affymetrix GeneChip S. cerevisiae tiling arrays	Granovskaia et al. (2010)
Ingolia_2009_canonical_uORFs_in_abundant_mRNAs	all upstream ORFs (uORFs) in highly transcribed 5' UTRs defined by their mRNA fragment densities being at least 100 rpkM with at least 50% of the mRNA densities of their associated protein-coding genes	Ingolia et al. (2009)
Ingolia_2009_canonical_uORFs	canonical upstream ORFs (uORFs) identified by finding AUG codons within annotated 5'UTRs and determination of the following open reading frame	Ingolia et al. (2009)

Ingolia_2009_noncanonical_translated_uORFs	non-AUG upstream ORFs (uORFs) in abundant mRNAs identified by finding codons with a single mismatch from AUG with an initiation context score of at least 0.01 within annotated 5'UTRs and determination of the following open reading frame	Ingolia et al. (2009)
Ingolia_2009_translated_uORFs	all uORFs with at least 1 rpM ribosome footprint occupancy	Ingolia et al. (2009)
Kertesz_2010_PARS_gene_annotations	global coordinates of transcribed genes for which structural profiles were obtained using the PARS method, including 3000 yeast coding transcripts, 14 tRNAs, 5 rRNAs, and 64 other annotated non-coding genes	Kertesz et al. (2010)
Lardenois_2011_noncoding_RNAs	1,452 differentially expressed ncRNAs identified using tiling arrays in this study	Lardenois et al. (2011)
Miura_2006_cDNA_clones	51,026 cDNA clones in two vector-capped cDNA libraries S288C-SD and SK1-Spo selected as"effective sequences" that have homology with both the 5'-flanking region of the cloning site of the vector (E-value >0.01) and the budding yeast genome sequence (E-value >10)	Miura et al. (2006)
Nagalakshmi_2008_3UTRs	predicted 3' UTR boundaries defined by RNA-seq	Nagalakshmi et al. (2008)
Nagalakshmi_2008_5UTRs	predicted 5' UTR boundaries defined by RNA-seq	Nagalakshmi et al. (2008)
Nagalakshmi_2008_novel_genes	novel gene predictions based on RNA-seq data	Nagalakshmi et al. (2008)
Nagalakshmi_2008_verified_introns	introns detected by a computational algorithm applied to the RNA-seq data	Nagalakshmi et al. (2008)
Neil_2009_class_I_CUTs	features that belong to clusters likely corresponding to"CUTs"	Neil et al. (2009)

Neil_2009_class_III_mRNAs	features that belong to clusters likely corresponding to“mRNA”	Neil et al. (2009)
Neil_2009_class_II_transcripts	features that belong to clusters corresponding to unclassified transcript features	Neil et al. (2009)
Neil_2009_class_I_ncRNAs	features that belong to clusters likely corresponding to“ncRNA”	Neil et al. (2009)
Neil_2009_class_I_other	features that belong to clusters likely corresponding to “other”	Neil et al. (2009)
Neil_2009_class_I_pre_mRNAs	features that belong to clusters likely corresponding to“pre-mRNA”	Neil et al. (2009)
van_Dijk_2011_XUTs	1658 Xrn1-sensitive unstable transcripts (XUTs)	van Dijk et al. (2011)
Waern_2013_AlphaFactor	differentially expressed genes under alpha factor treatment	Waern and Snyder (2013)
Waern_2013_Benomyl	differentially expressed genes under benomyl treatment	Waern and Snyder (2013)
Waern_2013_Calcofluor	differentially expressed genes under calcofluor treatment	Waern and Snyder (2013)
Waern_2013_CongoRed	differentially expressed genes under Congo red treatment	Waern and Snyder (2013)
Waern_2013_DNADamage	differentially expressed genes under methyl methanesulfonate treatment	Waern and Snyder (2013)
Waern_2013_GrapeJuice	differentially expressed genes under grape juice (Walgreens brand) treatment	Waern and Snyder (2013)
Waern_2013_HeatShock	differentially expressed genes under heat shock treatment	Waern and Snyder (2013)
Waern_2013_HighCalcium	differentially expressed genes under high calcium treatment	Waern and Snyder (2013)

Waern_2013_Hydroxyurea	differentially expressed genes under hydroxyurea treatment	Waern and Snyder (2013)
Waern_2013_LowNitrogen	differentially expressed genes under low nitrogen treatment	Waern and Snyder (2013)
Waern_2013_LowPhosphate	differentially expressed genes under low phosphate treatment	Waern and Snyder (2013)
Waern_2013_OxidativeStress	differentially expressed genes under hydrogen peroxide treatment	Waern and Snyder (2013)
Waern_2013_Salt	differentially expressed genes under sodium chloride treatment	Waern and Snyder (2013)
Waern_2013_ScGlycerolMedia	differentially expressed genes under glycerol synthetic complete medium treatment	Waern and Snyder (2013)
Waern_2013_ScMedia	differentially expressed genes under synthetic complete medium treatment	Waern and Snyder (2013)
Waern_2013_Sorbitol	differentially expressed genes under sorbitol treatment	Waern and Snyder (2013)
Waern_2013_StationaryPhase	differentially expressed genes under stationary phase treatment	Waern and Snyder (2013)
Xu_2009_CUTs	detected cryptic unstable transcripts (CUTs)	Xu et al. (2009)
Xu_2009_ORF-Ts	transcripts that overlapped with a verified or uncharacterized ORF as annotated in SGD	Xu et al. (2009)
Xu_2009_other_transcripts	transcripts that were not a) SUTs (no overlap with existing SGD annotation); b) ORF-Ts (overlapped with a verified or uncharacterized open reading frame in SGD); or c) CUTs (cryptic unstable transcripts)	Xu et al. (2009)
Xu_2009_SUTs	transcripts that did not overlap with existing annotation in SGD and exist stably in cells	Xu et al. (2009)

Yassour_2009_transcribed_regions	ab initio transcribed region predictions	Yassour et al. (2009)
Yassour_2009_UTR_boundaries	predicted 5' and 3' UTR boundaries	Yassour et al. (2009)
Yassour_2010_all_putative_antisense_transcripts	all identified transcribed units identified by the authors, from which a subset (1,103) were determined to be antisense, according to the following criterion: the unit must cover >=25% of any transcript from an opposite ORF	Yassour et al. (2010)
Yassour_2010_manual_antisense_transcripts	402 manually curated antisense units identified by the authors, a subset of the 1,103 antisense units listed in the"Yassour_2010_all_putative_antisense_transcript" track	Yassour et al. (2010)

Table A.2: Secondary Annotations with their descriptions and publications from which they were produced. Descriptions of Annotations are directly from SGD (Cherry et al., 2012). Each track is in GFF3 format unless specified (e.g. bedgraph). The superscripts next to track names indicate which category the track was classified under: 1 = DNA Damage, 2 = DNA-DNA Interactions, 3 = Histone Binding Sites, 4 = Modification or Tagging Sites, 5 = Other Binding Sites, 6 = Other Sequence Features, and 7 = Transcription Regulation.

Track Name	**Description of Annotation**	**Publication**
Albert_2007_H2AZ_nucleosome_positions [3]	distribution of budding yeast variant H2A.Z nucleosomes (unstranded)	Albert et al. (2007)
Albert_2007_H2AZ_nucleosome_positions_WC [3]	distribution of budding yeast variant H2A.Z nucleosomes (stranded)	Albert et al. (2007)
Buhler_2007_dmc1delta_DSB_hotspots_2x_threshold [1]	double strand break hotspots as determined from the dmc1delta mutant at a threshold 2x above the background	Buhler et al. (2007)
Buhler_2007_dmc1delta_DSB_hotspots_5x_threshold [1]	double strand break hotspots as determined from the dmc1delta mutant at a threshold 5x above the background	Buhler et al. (2007)
Buhler_2007_rad50S_DSB_hotspots_2x_threshold [1]	double strand break hotspots as determined from the rad50S mutant at a threshold 2x above the background	Buhler et al. (2007)
Buhler_2007_rad50S_DSB_hotspots_5x_threshold [1]	double strand break hotspots as determined from the rad50S mutant at a threshold 5x above the background	Buhler et al. (2007)

Buhler_2007_rad51delta_dmc1delta_DSB_hotspots_2x_threshold [1]	double strand break hotspots as determined from the rad51delta dmc1delta mutant at a threshold 2x above the background	Buhler et al. (2007)
Buhler_2007_rad51delta_dmc1delta_DSB_hotspots_5x_threshold [1]	double strand break hotspots as determined from the rad51delta dmc1delta mutant at a threshold 5x above the background	Buhler et al. (2007)
Eaton_2010_ORC_ACS [6]	253 likely ORC-binding locations associated ARS consensus sequences found by motif analysis on the ORC ChIP-seq peaks	Eaton et al. (2010)
Eaton_2010_orc1-161_mutant_G2_37C_mononucleosomal_fragments [3]	sequenced mononucleosome fragments following the digestion of isogenic orc1-161 temperature sensitive mutant with micrococcal nuclease at heat shock temperature (37C)	Eaton et al. (2010)
Eaton_2010_WT_async_23C_mononucleosomal_fragments [3]	sequenced mononucleosome fragments following the digestion of asynchronous WT cells with micrococcal nuclease at the permissive temperature (23C)	Eaton et al. (2010)
Eaton_2010_WT_G1_23C_mononucleosomal_fragments [3]	sequenced mononucleosome fragments following the digestion of G1 arrested WT cells with micrococcal nuclease at the permissive temperature (23C)	Eaton et al. (2010)
Eaton_2010_WT_G2_23C_mononucleosomal_fragments [3]	sequenced mononucleosome fragments following the digestion of G2 arrested WT cells with micrococcal nuclease at the permissive temperature (23C)	Eaton et al. (2010)

Field_2008_mapped_read_locations [3]	mapped sequence reads determined by BLASTing the raw reads to the yeast genome and retaining those that map uniquely to the genome, not overlapping with the rRNA locus, with lengths of 127-177bp and at least 95% identity	Field et al. (2008)
Hesselberth_2009_DNaseI_hypersensitive_sites [6]	protein-protected footprints detected at a q-value (false discovery rate) threshold of 0.1 by their assay	Hesselberth et al. (2009)
Lee_2007_HMM_nucleosome_state_calls [3]	well-positioned and fuzzy nucleosome calls made by a 78-state hidden-Markov model (HMM) trained on several loci where nucleosome positions are well-studied	Lee et al. (2007)
MacIsaac_2006_ChIP_chip_TFBSs [5]	binding sites determined from the ChIP-chip data of Harbison et al. (2004) at the most stringent binding p-value cutoff (0.001) and conservation cutoff studied	MacIsaac et al. (2006)
Mancera_2008_meiotic_recombination_hotspots [2]	total, crossover and non-crossover recombination hotspots	Mancera et al. (2008)
Mavrich_2008_H3H4_nucleosome_positions [3]	mapped nucleosome positions for each strand	Mavrich et al. (2008)
Mavrich_2008_H3H4_nucleosome_positions_WC [3]	mapped nucleosome positions for both strands	Mavrich et al. (2008)
Mayer_2010_Bur1_ChIP_chip [7]	normalized occupancy signals for transcription factor Bur1	Mayer et al. (2010)
Mayer_2010_Cet1_ChIP_chip [7]	normalized occupancy signals for transcription factor Cet1	Mayer et al. (2010)

Mayer_2010_Ctk1_ChIP_chip [7]	normalized occupancy signals for transcription factor Ctk1	Mayer et al. (2010)
Mayer_2010_Elf1_ChIP_chip [7]	normalized occupancy signals for transcription factor Elf1	Mayer et al. (2010)
Mayer_2010_Kin28_ChIP_chip [7]	normalized occupancy signals for transcription factor Kin28	Mayer et al. (2010)
Mayer_2010_Paf1_ChIP_chip [7]	normalized occupancy signals for transcription factor Paf1	Mayer et al. (2010)
Mayer_2010_Pcf11_ChIP_chip [7]	normalized occupancy signals for transcription factor Pcf11	Mayer et al. (2010)
Mayer_2010_Rpb1_CTD_Ser2P_ChIP_chip [7]	normalized occupancy signals for transcription factor Rpb1 phosphorylated at serine 2 residue of the C-terminal domain	Mayer et al. (2010)
Mayer_2010_Rpb1_CTD_Ser5P_ChIP_chip [7]	normalized occupancy signals for transcription factor Rpb1 phosphorylated at serine 5 residue of the C-terminal domain	Mayer et al. (2010)
Mayer_2010_Rpb1_CTD_Ser7P_ChIP_chip [7]	normalized occupancy signals for transcription factor Rpb1 phosphorylated at serine 7 residue of the C-terminal domain	Mayer et al. (2010)
Mayer_2010_Rpb3_ChIP_chip [7]	normalized occupancy signals for transcription factor Rpb3	Mayer et al. (2010)
Mayer_2010_Spn1_ChIP_chip [7]	normalized occupancy signals for transcription factor Spn1	Mayer et al. (2010)
Mayer_2010_Spt16_ChIP_chip [7]	normalized occupancy signals for transcription factor Spt16	Mayer et al. (2010)
Mayer_2010_Spt4_ChIP_chip [7]	normalized occupancy signals for transcription factor Spt4	Mayer et al. (2010)

Mayer_2010_Spt5_ChIP_chip [7]	normalized occupancy signals for transcription factor Spt5	Mayer et al. (2010)
Mayer_2010_Spt6_ChIP_chip [7]	normalized occupancy signals for transcription factor Spt6	Mayer et al. (2010)
Mayer_2010_Spt6deltaC_ChIP_chip [7]	normalized occupancy signals for transcription factor SptdeltaC	Mayer et al. (2010)
Mayer_2010_Tfg1_ChIP_chip [7]	normalized occupancy signals for transcription factor Tfg1	Mayer et al. (2010)
Mayer_2010_TFIIB_ChIP_chip [7]	normalized occupancy signals for transcription factor TFIIB	Mayer et al. (2010)
Ozsolak_2010_polyadenylation_sites [4]	polyadenylation sites as determined by clustering and grouping of the 5' ends of sequence reads produced by direct RNA sequencing (DRS)	Ozsolak et al. (2010)
Pan_2011_DSB_hotspots	double strand break hotspots [1]	Pan et al. (2011)
Pan_2011_Spo11_uniquely_mapped_reads [1]	sequence reads corresponding to Spo11-bound oligos mapping uniquely to the S288C genome	Pan et al. (2011)
Qi_2009_gene_conversions [2]	gene conversion events derived from 4 spores from a S288C/RM11 diploid	Qi et al. (2009)
Qi_2009_meiotic_crossovers [2]	91 meiotic crossover events derived from 4 spores from a S288C/RM11 diploid	Qi et al. (2009)
Rhee_2011_Gal4_ChIP_exo_bound_locations [5]	binding locations demarcated by the sequenced ChIP-exo peak pairs for the transcription factor Gal4	Rhee and Pugh (2011)
Rhee_2011_Phd1_ChIP_exo_bound_locations [5]	binding locations demarcated by the sequenced ChIP-exo peak pairs for the transcription factor Phd1	Rhee and Pugh (2011)

Rhee_2011_Rap1_ChIP_exo_bound_locations [5]	binding locations demarcated by the sequenced ChIP-exo peak pairs for the transcription factor Rap1	Rhee and Pugh (2011)
Rhee_2011_Reb1_ChIP_exo_bound_locations [5]	binding locations demarcated by the sequenced ChIP-exo peak pairs for the transcription factor Reb1	Rhee and Pugh (2011)
Rhee_2012_TATA_elements [6]	5,947 TATA-element containing features	Rhee and Pugh (2012)
Tachibana_2005_Adr1_and_Cat8_binding_regions [5]	ChIP-chip localization data for transcription factors Adr1 and Cat8 respectively	Tachibana et al. (2005)
Velculescu_1997_SAGE [4]	location of and number of occurrences of serial analysis of gene expression (SAGE) tags	Velculescu et al. (1997)
Venters_2011_all	genome-wide binding locations of 200 yeast transcription-related proteins under normal and acute heat-shock conditions by ChIP-chip	Venters et al. (2011)
Xu_2006_known_ARS_identified [6]	comparison of the authors' HMM predictions to 96 known ARS	Xu et al. (2006)
Xu_2006_MCM2_binding_regions [5]	281 MCM2-only binding regions detected by ChIP-chip	Xu et al. (2006)
Xu_2006_nimACS [6]	505 predicted ACS locations and sequences (nimACS) were determined from ORC and MCM2 ChIP-chip experiments	Xu et al. (2006)
Xu_2006_nimARS [6]	529 predicted ARS locations (nimARS) were determined from ORC and MCM2 ChIP-chip experiment	Xu et al. (2006)
Xu_2006_ORC_binding_regions [5]	47 ORC-only binding regions detected by ChIP-chip	Xu et al. (2006)

Xu_2006_ORC_MCM2_binding_regions [5]	383 (349 unique) ORC-MCM2 binding regions detected by ChIP-chip	Xu et al. (2006)
Zhang_2005_TSS [6]	transcription start sites	Zhang and Dietrich (2005)

Appendix B

Table B.1: The list of currently functional programs within the UAR-Pipeline, along with the options and descriptions for each program. See Table B.2 for descriptions of individual options. The Python packages called in the wrapper are sys (from the Python Standard Library) to append directories to the current working directory and the standard parser and standard logger modules written by Dr. Nick Schurch. The latter two modules were written to streamline the usage of the argparse, warnings, tempfile, and logging Python Standard Libraries together. General usage in a Linux/Unix environment is as follows: /sw/opt/python/2.7.3/bin/python /cluster/gjb_lab/ngiang/ workspace/GRNAseq/uar_pipeline/src/uar_pipeline.py program -option1 option1argument

Program	**Options**	**Description**
get-uar-gff3	-chrlengths -gtf -UARgff3	given the names and lengths of all chromosomes and a .gff3/.gtf file of genome annotations, produces a .gff3 file of UARs across the entire genome
wig-to-array	-chrlengths -wigsdir -prefix	given the names and lengths of all chromosomes and the directory where all RNA-seq alignment .wig files are (one per chromosome), produces a single numpy savez wigarrays.npz file in which values per base are stored as arrays (one per chromosome)

profile-uars	-wigarrays -gtf -prefix	given the numpy savez wigarrays.npz file and genome annotations .gff3/.gtf file, creates a pickled dictionary (see -uarprofile in Table B.2 for format)
uar-length-hist	-uarprofile -prefix	given a pickled dictionary of coordinates, lengths, and read counts for UARs, produces a histogram of the lengths of UARs
uar-totalreads-length	-uarprofile -prefix	given a pickled dictionary of coordinates, lengths, and read counts for UARs, produces a scatter plot of the number of total reads vs. the length of UARs
get-annot-source-feats	-gtf -prefix	given a .gff3/.gtf file of genome annotations, produces a tab-delimited list of unique source-feature pairs from the .gff3/.gtf file
uar-totalreads-2	-uarprofiles	given a list of two -uarprofiles (see -uarprofile in Table B.2), produces a scatter plot of total reads of first alignment vs. total reads of second alignment
uar-plotly-table	-uarprofiles -prefixes -tablename	given a list of -uarprofiles (see -uarprofile in Table B.2) and the names of each profile, creates a table (designated by -tablename) containing the columns chromosome, start, end, length of each UAR, along with the total reads from each alignment

ref-feat-plotly-table	-reffeatprofiles -prefixes -tablename	given a list of -reffeatprofiles (see Table B.2), creates a table (designated by -tablename) containing the columns chromosome, start, end, length of each annotation, along with the total reads from each alignment
sam-profile-uars	-chrlengths -sortedbam -gtf -prefix	given the names and lengths of chromosomes, a sorted .bam RNA-seq alignment file, and a .gff3/.gtf file of genome annotations, produces a pickled dictionary of read counts per base for all UARs
sam-profile-uar-orf-regions-text-output	-UARORFtab -sortedbam -prefix	given a tab-delimited list of all UAR ORFs (columns: chromosome, start, end, and frame) and a sorted .bam RNA-seq alignment file, produces a tab-delimited file of UARs (columns: chromosome, start, end, frame, id, and read depth)
sam-profile-regions-text-output	-regionsTAB -sortedbam -prefix	identical to sam-profile-uar-orf-regions-text-output, but without the frame column
bedgraph-pkl	-chrlengths -bgList -bgNames -prefix	given the names and lengths of chromosomes and a list of bedgraph files and their names, produces a pickled dictionary (see -bgpkl in Table B.2 for format)

uar-plotly-table-w-cons	-uarprofiles -prefixes -bgpkl -tablename	given a list of -uarprofiles (see -uarprofile in Table B.2), the names of the profiles, and a pickled dictionary with values from bedgraph file(s), produces a table (designated by -tablename) with the same information as uar-plotly-table. For each dictionary within the -bgpkl, the table also includes the sum of bedgraph values across the entire UAR. If the -bgpkl file includes a dictionary called 'multiz', the program will also find the maximum multiz score of any base for the UAR.

Table B.2: The list of options for programs within the UAR-Pipeline, along with a description and usage example for each. The list of programs that require these arguments are in Table B.1.

Option	**Description**	**Usage**
-bgList	list of bedgraph files	-bgList A.bedgraph B.bedgraph
-bgNames	list of bedgraph file names	-bgNames A B
-bgpkl	pickled dictionary containing scores per base from a bedgraph file (example of format for an artificial 8-bp chrI: 'chrI': 'A': numpy.array([0,0,4,0,2,2,2,2]), 'B': numpy.array([0,0,3,0,8,9,9,7]), 'chrII': ...)	-bgpkl bedgraphFile.pkl
-chrlengths	tab-delimited list of chromosome names and lengths	-chrlengths chrlengths.tab
-gtf	.gtf or .gff3 file containing genome annotations	-gtf annotations.gtf
-prefix	start of the output file names	-prefix near-default
-prefixes	list of starts of the output file names	-prefixes near-default unique stringent
-reffeatprofiles	list of pickled dictionaries containing read counts for a set of annotations (Format: ('SGD', 'gene'): 'chrI': 'A-total_reads': [28,24], 'B-total_reads': [18,14], 'length':[3,2],'pos':[[2,3,4],[3,4]], ('SGD', 'snoRNA'): ... . Two source-feature pairs are in tuples and are the dictionary keys (SGD gene and SGD snoRNA). All annotation entries are included (i.e. for SGD genes, there are two separate entries). For each entry, the total reads from each alignment of interest (A and B) are listed, as well as the length and positions/coordinates on the chromosome.)	-reffeatprofiles proteins.pkl snornas.pkl
-regionsTAB	tab-separated list of genomic regions of interest (columns: chromosome, start, end, id)	-regionsTAB genomic_regions.tab

-sortedbam	sorted and tabix'd .bam file of RNA-seq alignments	-sortedbam near-default_sorted.bam
-tablename	name of the table to be created	-tablename proteins
-UARgff3	path of the UAR .gff3 created	-UARgff3 uars.gff3
-UARORFtab	tab-separated list of ORFs (columns: chromosome, start, end, frame, id)	-UARORFtab uar_orfs.tab
-uarprofile	pickled dictionary containing coordinates, lengths, and read counts for UARs (format: 'chrI': 'pos': [[2,3,4],[7,8,9,10]], 'length': [3,4], 'total_reads': [15,24], 'chrII': ...)	-uarprofile near-default-SAM-uarProfile.pkl
-uarprofiles	list of uarprofiles (see -uarprofile)	-uarprofiles nd-SAM-uarProfile.pkl u-SAM-uarProfile.pkl
-wigarrays	numpy arrays in a single file in savez.npz format	-wigarrays wigarrays.npz
-wigsdir	directory where all .wig files are	-wigsdir /homes/ngiang/wigsdir/

Appendix C

Table C.1: The first 25 entries in BLASTX (v. 2.2.28+) results for chrI: 12,427–13,361. The database searched includes all non-redundant GenBank CDS translations, PDB, SwisProt, PIR, and PRF excluding environmental samples from WGS projects (Altschul et al., 1997).

Sequences producing significant alignments:	**Score (Bits)**	**E Value**
gb |EGA72474.1|Flo5p [Saccharomyces cerevisiae AWRI796]	77.8	7e-13
gb |EDZ71603.1|YHR211Wp-like protein [Saccharomyces cerevisia...]	77.8	7e-13
gb |EGA76195.1|Flo5p [Saccharomyces cerevisiae AWRI796]	74.7	4e-12
emb |CAY77581.1|Flo9p [Saccharomyces cerevisiae EC1118]	72.4	5e-11
dbj |BAG49462.1|flocculin [Saccharomyces pastorianus]	72.0	7e-11
dbj |BAA19915.1|flocculin [Saccharomyces cerevisiae]	71.2	1e-10
gb |AFH73975.1|flocculin protein FLO9 [Saccharomyces cerevisi...]	69.7	4e-10
gb |AFJ20718.1|flocculin [Saccharomyces cerevisiae]	69.3	6e-10
gb |AEC03967.1|FLO1 derivative [Saccharomyces cerevisiae/Schi...]	69.3	7e-10
ref |XP_001292657.1|hypothetical protein [Trichomonas vaginal...]	67.8	1e-09
gb |EGA80125.1|Flo5p [Saccharomyces cerevisiae Vin13]	67.0	2e-09
gb |EGA56258.1|Flo5p [Saccharomyces cerevisiae FostersB]	64.7	1e-08
emb |CDB64320.1|neurofilament [Clostridium clostridioforme CA...]	63.2	2e-08
gb |EGA84147.1|Flo5p [Saccharomyces cerevisiae Lalvin QA23]	59.3	6e-07
ref |XP_001623087.1|hypothetical protein NEMVEDRAFT_v1g219856...]	58.2	2e-06
gb |EGA84153.1|Flo9p [Saccharomyces cerevisiae VL3]	57.4	3e-06
gb |EMP33498.1|von Willebrand factor D and EGF domain-contain...	56.6	3e-06
ref |XP_001631735.1|predicted protein [Nematostella vectensis...]	51.6	7e-05
gb |EHN08658.1|Flo9p [Saccharomyces cerevisiae x Saccharomyce...]	49.7	2e-04
ref |XP_001279116.1|hypothetical protein [Trichomonas vaginal...]	50.1	2e-04
ref |XP_001280861.1|hypothetical protein [Trichomonas vaginal...]	48.5	6e-04
ref |XP_001282773.1|hypothetical protein [Trichomonas vaginal...]	50.1	7e-04
ref |XP_001297800.1|hypothetical protein [Trichomonas vaginal...]	50.1	0.001
ref |XP_001283617.1|hypothetical protein [Trichomonas vaginal...]	49.7	0.001
ref |XP_001295421.1|hypothetical protein [Trichomonas vaginal...]	49.7	0.001

Figure C.1: InterProScan (v. 4.8) results for the ORF at chrI: 11,569–13,174 showing Flocculin type 3 repeats as the only significant match.

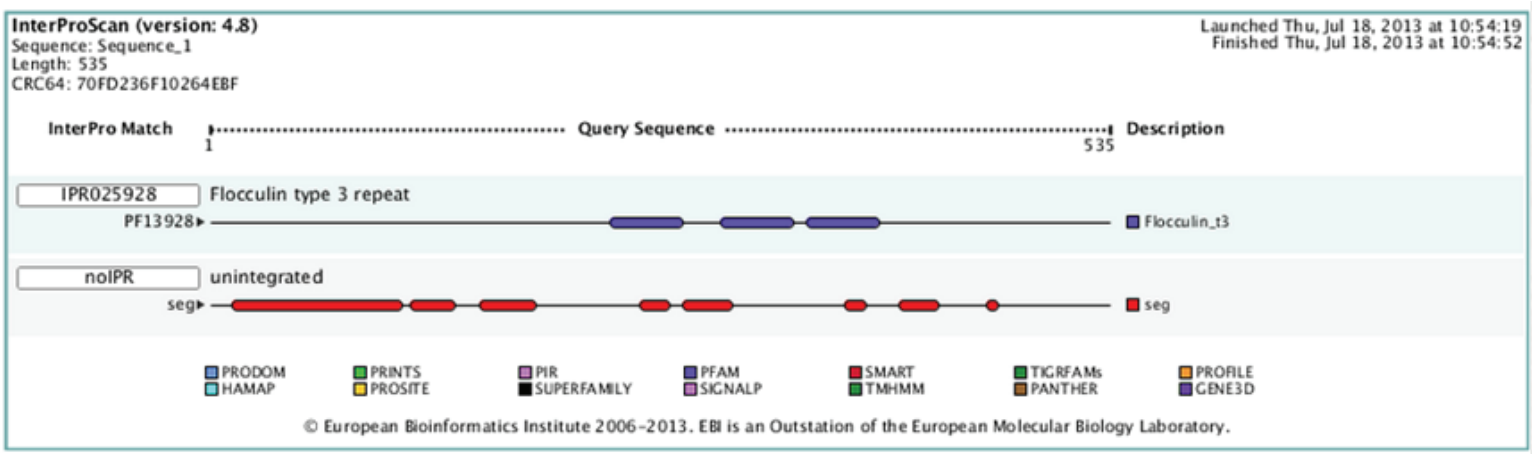


Figure C.2: InterPro results for FLO1, showing that signatures for the PA14 domain and Flocculin repeat were matched.

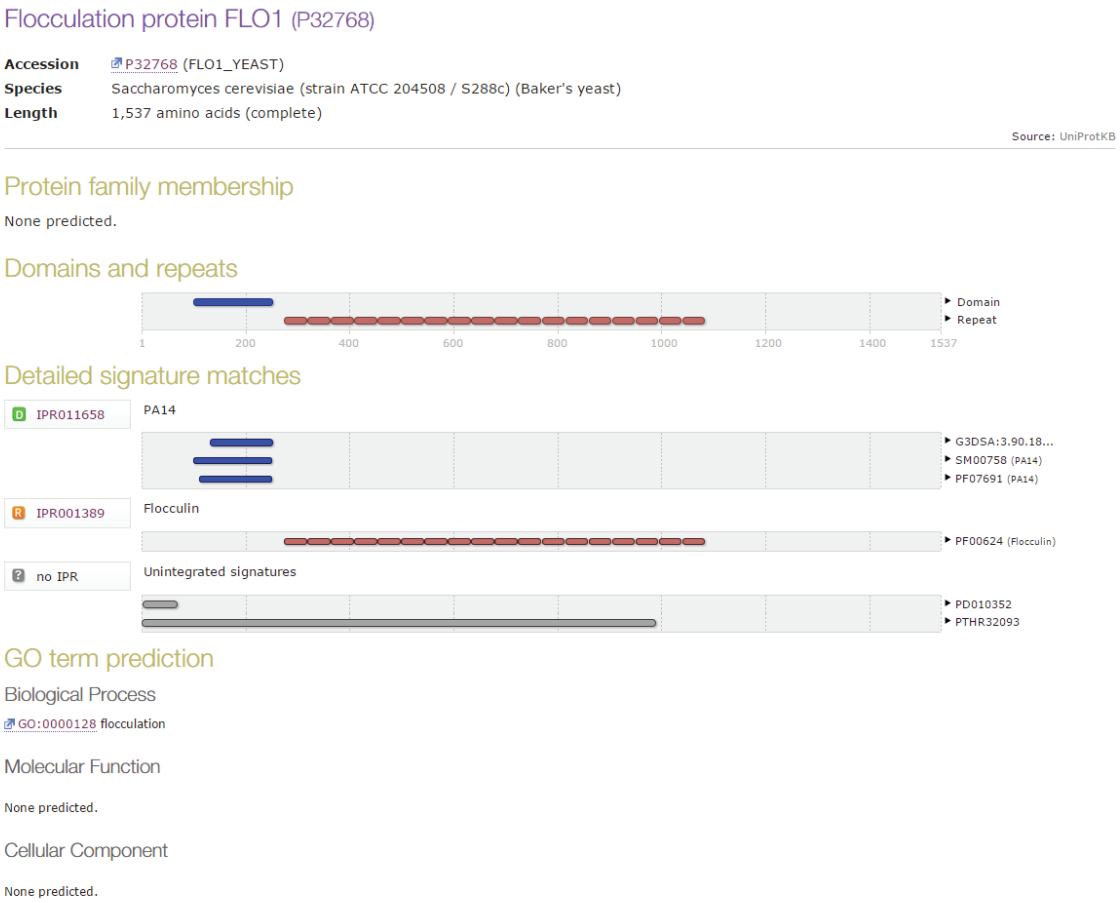


Figure C.3: InterPro results for FLO5, showing that signatures for the PA14 domain, Flocculin repeat, and Flocculin type 3 repeat were matched.

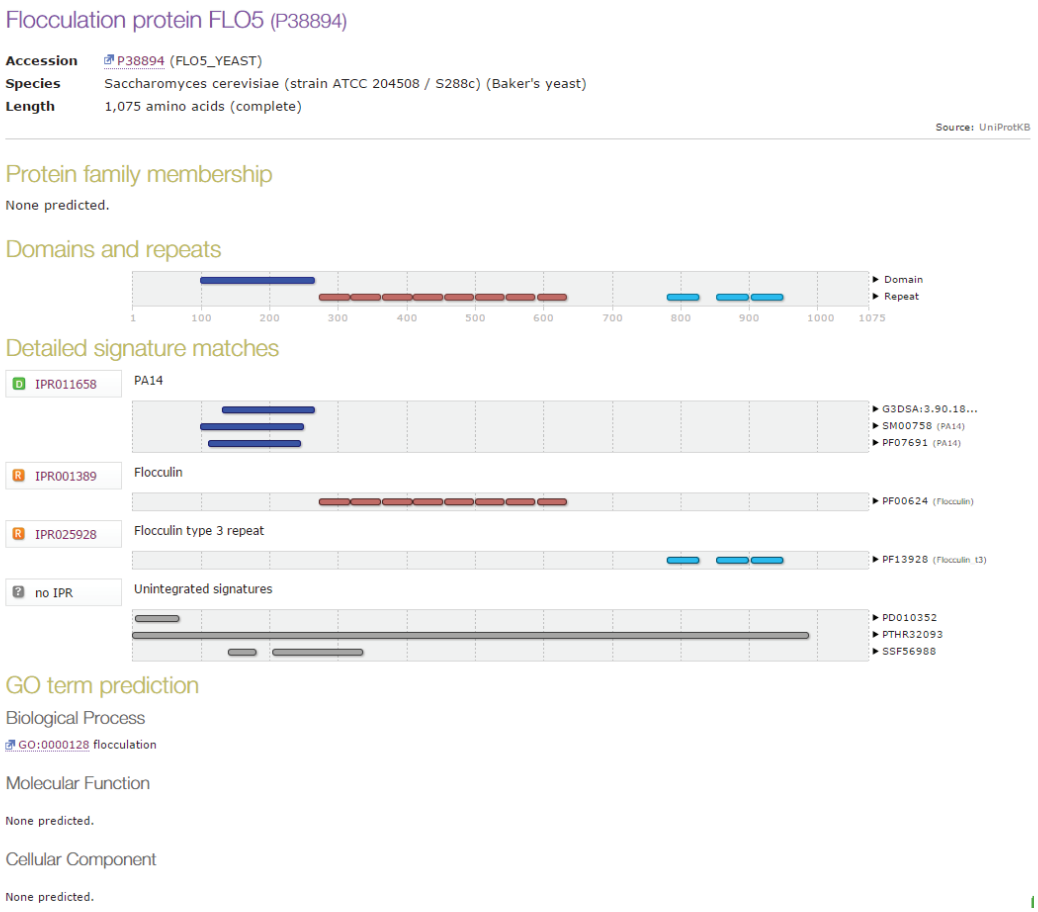


Figure C.4: InterPro results for FLO9, showing that signatures for the PA14 domain, Flocculin repeat, and Flocculin type 3 repeat were matched.

Figure C.5: Graphical summary of BLASTX results for the UAR chrV: 288,525–291,000, showing that the majority of hits align toward the 3' end of the region and have high alignment scores.

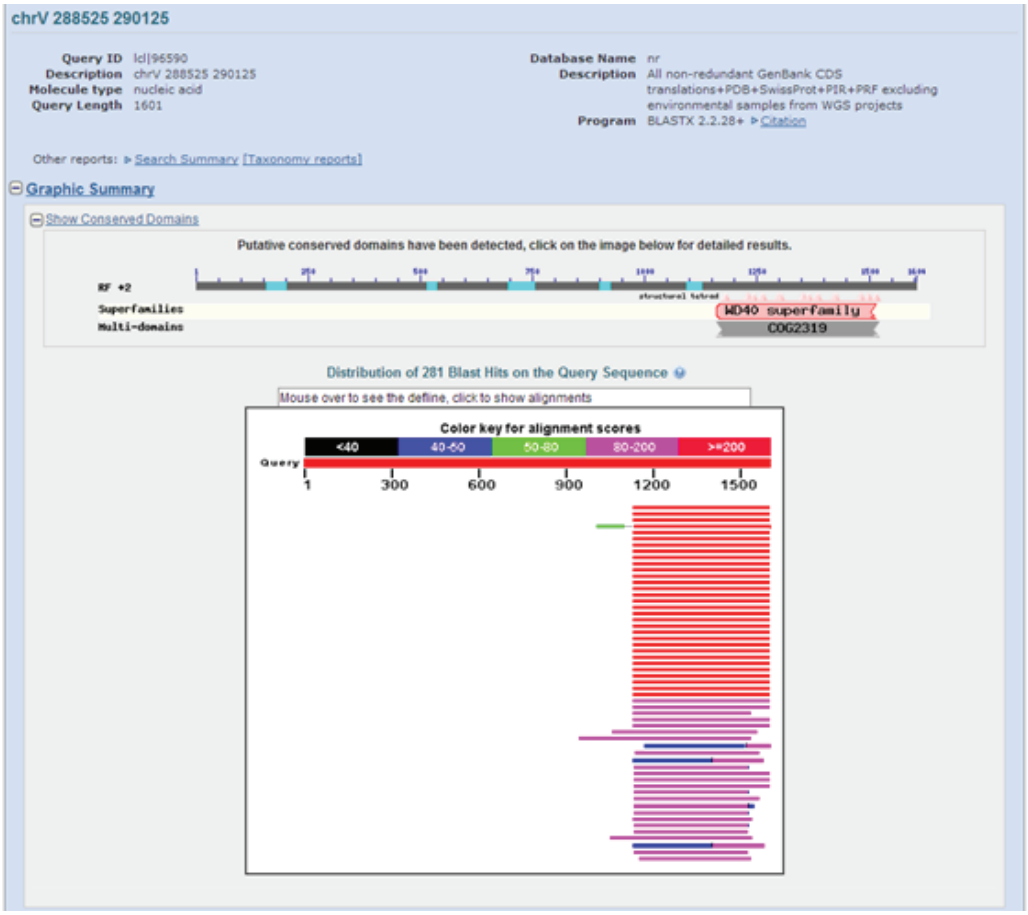


Figure C.6: Graphical summary of BLASTX results for the UAR chrV: 288,525–290,125, showing that the majority of hits span both ORFs at 289,528–289,905 and 289,908–290,799.

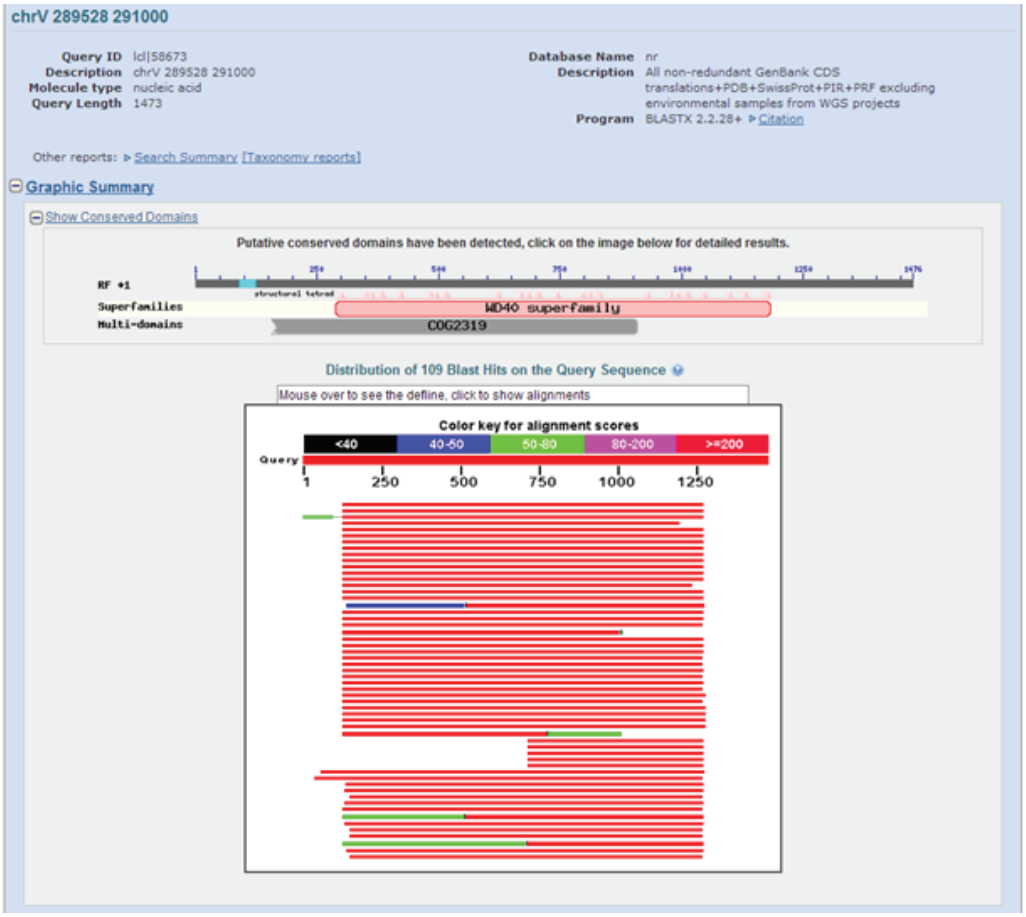
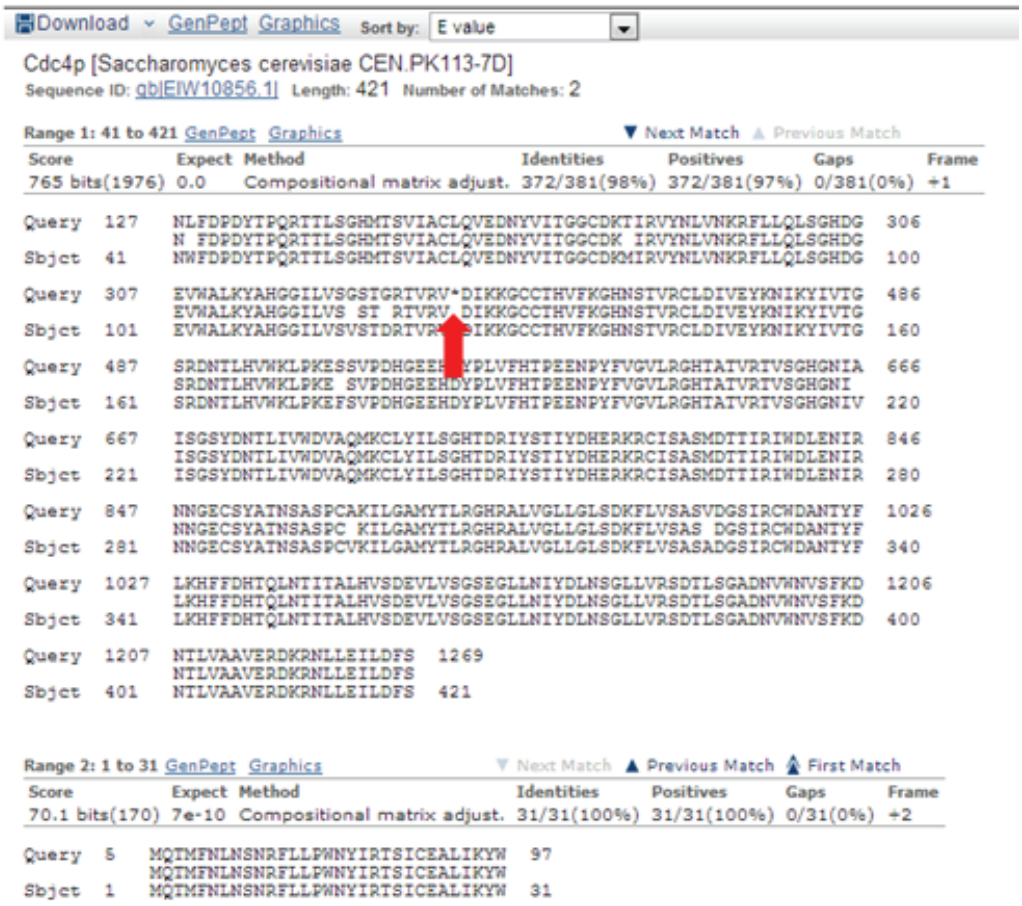


Figure C.7: This is an alignment showing an alignment of Cdc4 against the matched segment within the region chrV: 288,525–291,000. The red arrows indicates the stop codon between the two adjacent ORFs.



Download GenPept Graphics Sort by: E value

Cdc4p [Saccharomyces cerevisiae CEN.PK113-7D]

Sequence ID: gb|EIW10856.1| Length: 421 Number of Matches: 2

Range 1: 41 to 421 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
765 bits(1976)	0.0	Compositional matrix adjust.	372/381(98%)	372/381(97%)	0/381(0%)	+1

```
Query  127  NLFDPDYTPQRTTLSGHMTSVIACLQVEDNYVITGGCDKTIRVYNLVNKRFLLQLSGHDG  306
            N FDPDYTPQRTTLSGHMTSVIACLQVEDNYVITGGCDK IRVYNLVNKRFLLQLSGHDG
Sbjct  41   NWFDPDYTPQRTTLSGHMTSVIACLQVEDNYVITGGCDKMIRVYNLVNKRFLLQLSGHDG  100

Query  307  EVWALKYAHGGILVSGSTGRTVRV*DIKKGCCTHVFKGHNSTVRCLDIVEYKNIKYIVTG  486
            EVWALKYAHGGILVS ST RTVRV DIKKGCCTHVFKGHNSTVRCLDIVEYKNIKYIVTG
Sbjct  101  EVWALKYAHGGILVSVSTDRTVRV DIKKGCCTHVFKGHNSTVRCLDIVEYKNIKYIVTG  160

Query  487  SRDNTLHVWKLPKESSVPDHGEEHDYPLVFHTPEENPYFVGVLRGHTATVRTVSGHGNIA  666
            SRDNTLHVWKLPKE SVPDHGEEHDYPLVFHTPEENPYFVGVLRGHTATVRTVSGHGNI
Sbjct  161  SRDNTLHVWKLPKEFSVPDHGEEHDYPLVFHTPEENPYFVGVLRGHTATVRTVSGHGNIV  220

Query  667  ISGSYDNTLIVWDVAQMKCLYILSGHTDRIYSTIYDHERKRCISASMDTTIRIWDLENIR  846
            ISGSYDNTLIVWDVAQMKCLYILSGHTDRIYSTIYDHERKRCISASMDTTIRIWDLENIR
Sbjct  221  ISGSYDNTLIVWDVAQMKCLYILSGHTDRIYSTIYDHERKRCISASMDTTIRIWDLENIR  280

Query  847  NNGECSYATNSASPCAKILGAMYTLRGHRALVGLLGLSDKFLVSASVDGSIRCWDANTYF  1026
            NNGECSYATNSASPC KILGAMYTLRGHRALVGLLGLSDKFLVSAS DGSIRCWDANTYF
Sbjct  281  NNGECSYATNSASPCVKILGAMYTLRGHRALVGLLGLSDKFLVSASADGSIRCWDANTYF  340

Query  1027  LKHFFDHTQLNTITALHVSDEVLVSGSEGLLNIYDLNSGLLVRSDTLSGADNVWNVSFKD  1206
            LKHFFDHTQLNTITALHVSDEVLVSGSEGLLNIYDLNSGLLVRSDTLSGADNVWNVSFKD
Sbjct  341  LKHFFDHTQLNTITALHVSDEVLVSGSEGLLNIYDLNSGLLVRSDTLSGADNVWNVSFKD  400

Query  1207  NTLVAAVERDKRNLLEILDFS  1269
            NTLVAAVERDKRNLLEILDFS
Sbjct  401  NTLVAAVERDKRNLLEILDFS  421
```

Range 2: 1 to 31 GenPept Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
70.1 bits(170)	7e-10	Compositional matrix adjust.	31/31(100%)	31/31(100%)	0/31(0%)	+2

```
Query  5   MQTMFNLNSNRFLLPWNYIRTSICEALIKYW  97
           MQTMFNLNSNRFLLPWNYIRTSICEALIKYW
Sbjct  1   MQTMFNLNSNRFLLPWNYIRTSICEALIKYW  31
```

Appendix D

Table D.1: The 110 non-redundant masked SGD genes detected by the UAR Pipeline were searched against SGD's YeastMine database.

Gene Systematic Name	Gene Standard Name	Description
Q0140	VAR1	Mitochondrial ribosomal protein of the small subunit; mitochondrially-encoded; polymorphic in different strains due to variation in number of AAT (asparagine) codons; translated near the mitochondrial inner membrane; may have a role in loss of mitochondrial DNA under stress conditions
YAL003W	EFB1	Translation elongation factor 1 beta; stimulates nucleotide exchange to regenerate EF-1 alpha-GTP for the next elongation cycle; part of the EF-1 complex, which facilitates binding of aminoacyl-tRNA to the ribosomal A site; human homolog EEF1B2 can complement yeast efb1 mutants
YAL030W	SNC1	Vesicle membrane receptor protein (v-SNARE); involved in the fusion between Golgi-derived secretory vesicles with the plasma membrane; proposed to be involved in endocytosis; member of the synaptobrevin/VAMP family of R-type v-SNARE proteins; SNC1 has a paralog, SNC2, that arose from the whole genome duplication
YAL035W	FUN12	Translation initiation factor eIF5B; GTPase that promotes Met-tRNAiMet binding to ribosomes and ribosomal subunit joining; promotes GTP-dependent maturation of 18S rRNA by Nob1p; protein abundance increases in response to DNA replication stress; homolog of bacterial IF2
YBL019W	APN2	Class II abasic (AP) endonuclease involved in repair of DNA damage; homolog of human HAP1 and E. coli exoIII
YBL026W	LSM2	Lsm (Like Sm) protein; part of heteroheptameric complexes (Lsm2p-7p and either Lsm1p or 8p): cytoplasmic Lsm1p complex involved in mRNA decay; nuclear Lsm8p complex part of U6 snRNP and possibly involved in processing tRNA, snoRNA, and rRNA; relocalizes from nucleus to cytoplasmic foci upon DNA replication stress

YBL050W	SEC17	Alpha-SNAP cochaperone; SNARE-complex adaptor for Sec18 (NSF) during the disassembly of postfusion cis-SNARE complexes; stimulates the ATPase activity of Sec18p; peripheral membrane protein required for vesicular transport between ER and Golgi, the 'priming' step in homotypic vacuole fusion, and autophagy; similar to mammalian alpha-SNAP
YBL059W		Putative protein of unknown function; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies; YBL059W has a paralog, YER093C-A, that arose from the whole genome duplication
YBL081W		Non-essential protein of unknown function; null mutation results in a decrease in plasma membrane electron transport
YBL084C	CDC27	Subunit of the Anaphase-Promoting Complex/Cyclosome (APC/C); APC/C is a ubiquitin-protein ligase required for degradation of anaphase inhibitors, including mitotic cyclins, during the metaphase/anaphase transition
YBL092W	RPL32	Ribosomal 60S subunit protein L32; overexpression disrupts telomeric silencing; homologous to mammalian ribosomal protein L32, no bacterial homolog
YBR108W	AIM3	Protein that inhibits barbed-end actin filament elongation; interacts with Rvs167p; null mutant is viable and displays elevated frequency of mitochondrial genome loss
YBR158W	AMN1	Protein required for daughter cell separation; multiple mitotic checkpoints, and chromosome stability; contains 12 degenerate leucine-rich repeat motifs; expression is induced by the Mitotic Exit Network (MEN)
YBR161W	CSH1	Mannosylinositol phosphorylceramide (MIPC) synthase catalytic subunit; forms a complex with regulatory subunit Csg2p; function in sphingolipid biosynthesis is overlapping with that of Sur1p; CSH1 has a paralog, SUR1, that arose from the whole genome duplication
YBR189W	RPS9B	Protein component of the small (40S) ribosomal subunit; homologous to mammalian ribosomal protein S9 and bacterial S4; RPS9B has a paralog, RPS9A, that arose from the whole genome duplication
YBR191W	RPL21A	Ribosomal 60S subunit protein L21A; homologous to mammalian ribosomal protein L21, no bacterial homolog; RPL21A has a paralog, RPL21B, that arose from the whole genome duplication

YBR212W	NGR1	RNA binding protein that negatively regulates growth rate; interacts with the 3' UTR of the mitochondrial porin (POR1) mRNA and enhances its degradation; overexpression impairs mitochondrial function; interacts with Dhh1p to mediate POR1 mRNA decay; expressed in stationary phase
YCL005W-A	VMA9	Vacuolar H+ ATPase subunit e of the V-ATPase V0 subcomplex; essential for vacuolar acidification; interacts with the V-ATPase assembly factor Vma21p in the ER; involved in V0 biogenesis
YDL012C		Tail-anchored plasma membrane protein with a conserved CYSTM module; possibly involved in response to stress; may contribute to non-homologous end-joining (NHEJ) based on ydl012c htz1 double null phenotype; YDL012C has a paralog, YBR016W, that arose from the whole genome duplication
YDL029W	ARP2	Essential component of the Arp2/3 complex; Arp2/3 is a highly conserved actin nucleation center required for the motility and integrity of actin patches; involved in endocytosis and membrane growth and polarity; required for efficient Golgi-to-ER trafficking in COPI mutants
YDL058W	USO1	Essential protein involved in vesicle-mediated ER to Golgi transport; binds membranes and functions during vesicle docking to the Golgi; required for assembly of the ER-to-Golgi SNARE complex
YDL064W	UBC9	SUMO-conjugating enzyme involved in the Smt3p conjugation pathway; nuclear protein required for S- and M-phase cyclin degradation and mitotic control; involved in proteolysis mediated by the anaphase-promoting complex cyclosome (APCC)
YDL075W	RPL31A	Ribosomal 60S subunit protein L31A; associates with karyopherin Sxm1p; loss of both Rpl31p and Rpl39p confers lethality; homologous to mammalian ribosomal protein L31, no bacterial homolog; RPL31A has a paralog, RPL31B, that arose from the whole genome duplication
YDL108W	KIN28	Serine/threonine protein kinase, subunit of transcription factor TFIIH; involved in transcription initiation at RNA polymerase II promoters; phosphorylates Ser5 residue of the PolII C-terminal domain (CTD) at gene promoters; relocalizes to the cytosol in response to hypoxia
YDL130W	RPP1B	Ribosomal protein P1 beta; component of the ribosomal stalk, which is involved in interaction of translational elongation factors with ribosome; free (non-ribosomal) P1 stimulates the phosphorylation of the eIF2 alpha subunit (Sui2p) by Gcn2p; accumulation is regulated by phosphorylation and interaction with the P2 stalk component

YDL189W	RBS1	Protein involved in assembly of the RNA polymerase III (Pol III) complex; high copy suppressor of Pol III assembly mutation and psk1 psk2 mutations that confer temperature-sensitivity for galactose utilization; physically interacts with Pol III; proposed to bind single-stranded nucleic acids via its R3H domain
YDL219W	DTD1	D-Tyr-tRNA(Tyr) deacylase; functions in protein translation, may affect nonsense suppression via alteration of the protein synthesis machinery; ubiquitous among eukaryotes
YDR064W	RPS13	Protein component of the small (40S) ribosomal subunit; homologous to mammalian ribosomal protein S13 and bacterial S15
YDR074W	TPS2	Phosphatase subunit of the trehalose-6-P synthase/phosphatase complex; involved in synthesis of the storage carbohydrate trehalose; expression is induced by stress conditions and repressed by the Ras-cAMP pathway; protein abundance increases in response to DNA replication stress
YDR092W	UBC13	E2 ubiquitin-conjugating enzyme; involved in the error-free DNA postreplication repair pathway; interacts with Mms2p to assemble ubiquitin chains at the Ub Lys-63 residue; DNA damage triggers redistribution from the cytoplasm to the nucleus
YDR099W	BMH2	14-3-3 protein, minor isoform; controls proteome at post-transcriptional level, binds proteins and DNA, involved in regulation of many processes including exocytosis, vesicle transport, Ras/MAPK signaling, and rapamycin-sensitive signaling; protein increases in abundance and relative distribution to the nucleus increases upon DNA replication stress; abundance relative to Bmh1p increases during sporulation
YDR228C	PCF11	mRNA 3' end processing factor; essential component of cleavage and polyadenylation factor IA (CF IA), involved in pre-mRNA 3' end processing and in transcription termination; binds C-terminal domain of largest subunit of RNA pol II (Rpo21p); required for gene looping; relocalizes to the cytosol in response to hypoxia
YDR381W	YRA1	Nuclear polyadenylated RNA-binding protein; required for export of poly(A)+ mRNA from the nucleus; proposed to couple mRNA export with 3' end processing via its interactions with Mex67p and Pcf11p; interacts with DBP2; inhibits the helicase activity of Dbp2; functionally redundant with Yra2p, another REF family member

YDR390C	UBA2	Subunit of heterodimeric nuclear SUMO activating enzyme E1 with Aos1p; activates Smt3p (SUMO) before its conjugation to proteins (sumoylation), which may play a role in protein targeting; essential for viability
YDR471W	RPL27B	Ribosomal 60S subunit protein L27B; homologous to mammalian ribosomal protein L27, no bacterial homolog; RPL27B has a paralog, RPL27A, that arose from the whole genome duplication
YEL003W	GIM4	Subunit of the heterohexameric cochaperone prefoldin complex; complex binds specifically to cytosolic chaperonin and transfers target proteins to it
YEL012W	UBC8	Ubiquitin-conjugating enzyme that regulates gluconeogenesis; negatively regulates gluconeogenesis by mediating the glucose-induced ubiquitination of fructose-1,6-bisphosphatase (FBPase); cytoplasmic enzyme that catalyzes the ubiquitination of histones in vitro
YER129W	SAK1	Upstream serine/threonine kinase for the SNF1 complex; plays a role in pseudohyphal groth; partially redundant with Elm1p and Tos3p; members of this family have functional orthology with LKB1, a mammalian kinase associated with Peutz-Jeghers cancer-susceptibility syndrome; SAK1 has a paralog, TOS3, that arose from the whole genome duplication
YER133W	GLC7	Type 1 S/T protein phosphatase (PP1) catalytic subunit; involved in glycogen metabolism, sporulation and mitotic progression; interacts with multiple regulatory subunits; regulates actomyosin ring formation; subunit of CPF; recruited to mating projections by Afr1p interaction; regulates nucleocytoplasmic shuttling of Hxk2p; import into the nucleus is inhibited during spindle assembly checkpoint arrest; involved in dephosphorylating Rps6a/b and Bnr1p
YFL033C	RIM15	Protein kinase involved in cell proliferation in response to nutrients; glucose-repressible; involved in signal transduction during cell proliferation in response to nutrients, specifically the establishment of stationary phase; identified as a regulator of IME2; phosphorylates Igo1p and Igo2p; substrate of Pho80p-Pho85p kinase
YFR045W		Putative mitochondrial transport protein; null mutant is viable, exhibits decreased levels of chitin and normal resistance to calcofluor white
YGL028C	SCW11	Cell wall protein with similarity to glucanases; may play a role in conjugation during mating based on its regulation by Ste12p

YGL030W	RPL30	Ribosomal 60S subunit protein L30; involved in pre-rRNA processing in the nucleolus; autoregulates splicing of its transcript; homologous to mammalian ribosomal protein L30, no bacterial homolog
YGL066W	SGF73	Subunit of DUBm module of SAGA and SLIK; has roles in anchoring deubiquitination module (DUBm) into SAGA and SLIK complexes, maintaining organization and ubiquitin-binding conformation of Ubp8p, thereby contributing to overall DUBm activity; involved in preinitiation complex assembly at promoters; relocalizes to cytosol under hypoxia; human homolog ATXN7 implicated in spinocerebellar ataxia, and can complement yeast null mutant
YGL103W	RPL28	Ribosomal 60S subunit protein L28; homologous to mammalian ribosomal protein L27A and bacterial L15; may have peptidyl transferase activity; can mutate to cycloheximide resistance
YGL131C	SNT2	Subunit of Snt2C complex, RING finger ubiquitin ligase (E3); physically associates with Ecm5p and Rpd3p; along with Ecm5p, recruits Rpd3p to small number of promoters; colocalizes with Ecm5p, independently of Rpd3p, to promoters of stress response genes upon oxidative stress; involved in ubiquitination, degradation of excess histones; interacts with Ubc4p; role in regulating genes encoding amine transporters; relocalizes from nucleus to cytoplasm upon DNA replication stress
YGL137W	SEC27	Essential beta'-coat protein of the COPI coatomer; involved in ER-to-Golgi and Golgi-to-ER transport; contains WD40 domains that mediate cargo selective interactions; 45% sequence identity to mammalian beta'-COP
YGL232W	TAN1	Putative tRNA acetyltransferase; RNA-binding protein required for the formation of the modified nucleoside N(4)-acetylcytidine in serine and leucine tRNAs but not required for the same modification in 18S rRNA; protein abundance increases in response to DNA replication stress
YGL251C	HFM1	Meiosis specific DNA helicase; involved in the conversion of double-stranded breaks to later recombination intermediates and in crossover control; catalyzes the unwinding of Holliday junctions; has ssDNA and dsDNA stimulated ATPase activity
YGR029W	ERV1	Flavin-linked sulfhydryl oxidase of the mitochondrial IMS; N-terminus is an intrinsically disordered domain that in the cytosol helps target Erv1p to mitochondria, and in the intermembrane space oxidizes Mia40p as part of a disulfide relay system that promotes intermembrane space retention of imported proteins; functional ortholog of human GFER (ALR); human GFER carrying N-terminal 21 amino acids of Erv1p functionally complements the lethality of the erv1 null mutation

YGR034W	RPL26B	Ribosomal 60S subunit protein L26B; binds to 5.8S rRNA; non-essential even when paralog is also deleted; deletion has minimal affections on ribosome biosynthesis; homologous to mammalian ribosomal protein L26 and bacterial L24; RPL26B has a paralog, RPL26A, that arose from the whole genome duplication
YGR238C	KEL2	Protein that negatively regulates mitotic exit; forms a complex with Kel1p and Bud14p that regulates Bnr1p (formin) to affect actin cable assembly, cytokinesis, and polarized growth; functions in a complex with Kel1p, interacts with Tem1p and Lte1p; localizes to regions of polarized growth; potential Cdc28p substrate
YHL001W	RPL14B	Ribosomal 60S subunit protein L14B; homologous to mammalian ribosomal protein L14, no bacterial homolog; RPL14B has a paralog, RPL14A, that arose from the whole genome duplication; protein abundance increases in response to DNA replication stress
YHR012W	VPS29	Subunit of the membrane-associated retromer complex; endosomal protein; essential for endosome-to-Golgi retrograde transport; forms a subcomplex with Vps35p and Vps26p that selects cargo proteins for endosome-to-Golgi retrieval
YHR030C	SLT2	Serine/threonine MAP kinase; coordinates expression of all 19S regulatory particle assembly-chaperones (RACs) to control proteasome abundance; involved in regulating maintenance of cell wall integrity, cell cycle progression, nuclear mRNA retention in heat shock, septum assembly; required for mitophagy, pexophagy; affects recruitment of mitochondria to phagophore assembly site; plays role in adaptive response of cells to cold; regulated by the PKC1-mediated signaling pathway
YHR165C	PRP8	Component of U4/U6-U5 snRNP complex; involved in second catalytic step of splicing; participates in spliceosomal assembly through its interaction with U1 snRNA; largest and most evolutionarily conserved protein of the spliceosome; mutations in human ortholog, PRPF8, cause Retinitis pigmentosa and missplicing in Myelodysplastic syndrome; mouse ortholog interacts with androgen receptor and may have a role in prostate cancer
YIL106W	MOB1	Component of the mitotic exit network; associates with and is required for the activation and Cdc15p-dependent phosphorylation of the Dbf2p kinase; required for cytokinesis and cell separation; component of the CCR4 transcriptional complex; relocalizes from cytoplasm to the nuclear periphery upon DNA replication stress

YIL156W	UBP7	Ubiquitin-specific protease that cleaves ubiquitin-protein fusions; UBP7 has a paralog, UBP11, that arose from the whole genome duplication
YIL156W-B		Putative protein of unknown function; originally identified based on homology to gossypii</i>and other related yeasts; SWAT-GFP and mCherry fusion proteins localize to the vacuole, while SWAT-GFP fusion also localizes to the endoplasmic reticulum
YJL001W	PRE3	Beta 1 subunit of the 20S proteasome; responsible for cleavage after acidic residues in peptides
YJL008C	CCT8	Subunit of the cytosolic chaperonin Cct ring complex; related to Tcp1p, required for the assembly of actin and tubulins in vivo
YKL006W	RPL14A	Ribosomal 60S subunit protein L14A; N-terminally acetylated; homologous to mammalian ribosomal protein L14, no bacterial homolog; RPL14A has a paralog, RPL14B, that arose from the whole genome duplication
YKL032C	IXR1	Transcriptional repressor that regulates hypoxic genes during normoxia; involved in the aerobic repression of genes such as COX5b, TIR1, and HEM13; binds DNA intrastrand cross-links formed by cisplatin; HMG (high mobility group box) domain containing protein which binds and bends cisplatin-modified DNA, blocking excision repair; IXR1 has a paralog, ABF2, that arose from the whole genome duplication
YKL048C	ELM1	Serine/threonine protein kinase; regulates the orientation checkpoint, the morphogenesis checkpoint and the metabolic switch from fermentative to oxidative metabolism by phosphorylating the activation loop of Kin4p, Hsl1p and Snf4p respectively; cooperates with Hsl7p in recruiting Hsl1p to the septin ring, a prerequisite for subsequent recruitment, phosphorylation, and degradation of Swe1p; forms part of the bud neck ring; regulates cytokinesis
YKL081W	TEF4	Gamma subunit of translational elongation factor eEF1B; stimulates the binding of aminoacyl-tRNA (AA-tRNA) to ribosomes by releasing eEF1A (Tef1p/Tef2p) from the ribosomal complex
YKL134C	Oct1	Mitochondrial intermediate peptidase; cleaves destabilizing N-terminal residues of a subset of proteins upon import, after their cleavage by mitochondrial processing peptidase (Mas1p-Mas2p); may contribute to mitochondrial iron homeostasis
YKL150W	MCR1	Mitochondrial NADH-cytochrome b5 reductase; involved in ergosterol biosynthesis

YKL156W	RPS27A	Protein component of the small (40S) ribosomal subunit; homologous to mammalian ribosomal protein S27, no bacterial homolog; RPS27A has a paralog, RPS27B, that arose from the whole genome duplication; protein abundance increases in response to DNA replication stress
YKL157W	APE2	Aminopeptidase yscII; may have a role in obtaining leucine from dipeptide substrates; APE2 has a paralog, AAP1, that arose from the whole genome duplication
YKL190W	CNB1	Calcineurin B; regulatory subunit of calcineurin, a Ca++/calmodulin-regulated type 2B protein phosphatase which regulates Crz1p (stress-response transcription factor); other calcineurin subunit encoded by CNA1 and/or CMP1; regulates function of Aly1p alpha-arrestin; myristoylation by Nmt1p reduces calcineurin activity in response to submaximal Ca signals, is needed to prevent constitutive phosphatase activity; protein abundance increases in response to DNA replication stress
YKR057W	RPS21A	Protein component of the small (40S) ribosomal subunit; homologous to mammalian ribosomal protein S21, no bacterial homolog; RPS21A has a paralog, RPS21B, that arose from the whole genome duplication
YKR095W-A	PCC1	Component of the EKC/KEOPS protein complex; EKC/KEOPS complex is required for t6A tRNA modification and telomeric TG1-3 recombination; may have role in transcription; other complex members are Kae1p, Gon7p, Bud32p, and Cgi121p
YLL019C	KNS1	Protein kinase involved in negative regulation of PolIII transcription; effector kinase of the TOR signaling pathway and phosphorylates Rpc53p to regulate ribosome and tRNA biosynthesis; member of the LAMMER family of protein kinases, which are serine/threonine kinases also capable of phosphorylating tyrosine residues; capable of autophosphorylation
YLR061W	RPL22A	Ribosomal 60S subunit protein L22A; required for translation of long 5' UTR of IME1 mRNA and meiotic entry; required for the oxidative stress response, pseudohyphal and invasive growth; homologous to mammalian ribosomal protein L22, no bacterial homolog; RPL22A has a paralog, RPL22B, that arose from the whole genome duplication
YLR185W	RPL37A	Ribosomal 60S subunit protein L37A; required for processing of 27SB pre-rRNA and formation of stable 66S assembly intermediates; homologous to mammalian ribosomal protein L37, no bacterial homolog; RPL37A has a paralog, RPL37B, that arose from the whole genome duplication

YLR256W	HAP1	Zinc finger transcription factor; involved in the complex regulation of gene expression in response to levels of heme and oxygen; localizes to the mitochondrion as well as to the nucleus; the S288C sequence differs from other strain backgrounds due to a Ty1 insertion in the carboxy terminus
YLR283W		Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to mitochondria; YLR283W is not an essential gene
YLR383W	SMC6	Component of the SMC5-SMC6 complex; this complex plays a key role in the removal of X-shaped DNA structures that arise between sister chromatids during DNA replication and repair; homologous to S. pombe rad18
YLR448W	RPL6B	Ribosomal 60S subunit protein L6B; binds 5.8S rRNA; homologous to mammalian ribosomal protein L6, no bacterial homolog; RPL6B has a paralog, RPL6A, that arose from the whole genome duplication
YML036W	CGI121	Component of the EKC/KEOPS complex; EKC/KEOPS complex is required for t6A tRNA modification and telomeric TG1-3 recombination; may have role in transcription; Cgi121p is dispensable for tRNA modification; other complex members are Bud32p, Kae1p, Pcc1p, and Gon7p
YML066C	SMA2	Meiosis-specific prospore membrane protein; required to produce bending force necessary for proper assembly of the prospore membrane during sporulation
YML094W	GIM5	Subunit of the heterohexameric cochaperone prefoldin complex; prefoldin binds specifically to cytosolic chaperonin and transfers target proteins to it; prefoldin complex also localizes to chromatin of actively transcribed genes in the nucleus and facilitates transcriptional elongation
YMR070W	MOT3	Transcriptional repressor, activator; role in cellular adjustment to osmotic stress including modulation of mating efficiency; involved in repression of subset of hypoxic genes by Rox1p, repression of several DAN/TIR genes during aerobic growth, ergosterol biosynthetic genes in response to hyperosmotic stress; contributes to recruitment of Tup1p-Cyc8p general repressor to promoters; relocalizes to cytosol under hypoxia; forms [MOT3+] prion under anaerobic conditions
YMR079W	SEC14	Phosphatidylinositol/phosphatidylcholine transfer protein; involved in regulating PtdIns, PtdCho, and ceramide metabolism, products of which regulate intracellular transport and UPR; has a role in localization of lipid raft proteins; functionally homologous to mammalian PITPs; SEC14 has a paralog, YKL091C, that arose from the whole genome duplication

YMR135C	GID8	Subunit of GID Complex, binds strongly to central component Vid30p; GID Complex is involved in proteasome-dependent catabolite inactivation of fructose-1,6-bisphosphatase; recruits Rmd5p, Fyv10 and Vid28p to GID Complex; contains LisH, CTLH, and CRA domains that mediate binding to Vid30p (LisH) and Rmd5p and Vid28p (CTLH and CRA); dosage-dependent regulator of START
YMR194W	RPL36A	Ribosomal 60S subunit protein L36A; N-terminally acetylated; binds to 5.8 S rRNA; homologous to mammalian ribosomal protein L36, no bacterial homolog; RPL36A has a paralog, RPL36B, that arose from the whole genome duplication
YNL044W	YIP3	Protein localized to COPII vesicles; proposed to be involved in ER to Golgi transport; interacts with members of the Rab GTPase family and Yip1p; also interacts with Rtn1p
YNL112W	DBP2	ATP-dependent RNA helicase of the DEAD-box protein family; has strong preference for dsRNA; interacts with YRA1; required for assembly of Yra1p, Nab2p and Mex67p onto mRNA and formation of nuclear mRNP; involved in mRNA decay and rRNA processing; may be involved in suppression of transcription from cryptic initiation sites
YNL138W-A	YSF3	Component of the SF3b subcomplex of the U2 snRNP; essential protein required for splicing and for assembly of SF3b
YNL243W	SLA2	Adaptor protein that links actin to clathrin and endocytosis; involved in membrane cytoskeleton assembly and cell polarization; present in the actin cortical patch of the emerging bud tip; dimer in vivo
YNL246W	VPS75	NAP family histone chaperone; binds to histones and Rtt109p, stimulating histone acetyltransferase activity; possesses nucleosome assembly activity in vitro; proposed role in vacuolar protein sorting and in double-strand break repair; protein abundance increases in response to DNA replication stress; relocalizes to the cytosol in response to hypoxia
YNL327W	EGT2	Glycosylphosphatidylinositol (GPI)-anchored cell wall endoglucanase; required for proper cell separation after cytokinesis; expression is activated by Swi5p and tightly regulated in a cell cycle-dependent manner
YOR060C	SLD7	Protein with a role in chromosomal DNA replication; interacts with Sld3p and reduces its affinity for Cdc45p; deletion mutant has aberrant mitochondria

YOR096W	RPS7A	Protein component of the small (40S) ribosomal subunit; interacts with Kti11p; deletion causes hypersensitivity to zymocin; homologous to mammalian ribosomal protein S7, no bacterial homolog; RPS7A has a paralog, RPS7B, that arose from the whole genome duplication
YOR239W	ABP140	AdoMet-dependent tRNA methyltransferase and actin binding protein; C-terminal domain is responsible for 3-methylcytidine modification of residue 32 of the tRNA anticodon loop of tRNA-Thr and tRNA-Ser and contains an S-adenosylmethionine (AdoMet) binding motif; N-terminal actin binding sequence interacts with actin filaments and localizes to actin patches and cables; N- and C-terminal domains are encoded in separate ORFs that are translated into one protein via a +1 frameshift
YOR291W	YPK9	Vacuolar protein with a possible role in sequestering heavy metals; has similarity to the type V P-type ATPase Spf1p; homolog of human ATP13A2 (PARK9), mutations in which are associated with Parkinson disease and Kufor-Rakeb syndrome
YOR296W		Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm; expressed during copper starvation; YOR296W is not an essential gene
YPL016W	SWI1	Subunit of the SWI/SNF chromatin remodeling complex; regulates transcription by remodeling chromatin; required for transcription of many genes, including ADH1, ADH2, GAL1, HO, INO1 and SUC2; self-assembles to form [SWI+] prion and to alter expression pattern; human homolog ARID1A is a candidate tumor suppressor gene in breast cancer
YPL079W	RPL21B	Ribosomal 60S subunit protein L21B; homologous to mammalian ribosomal protein L21, no bacterial homolog; RPL21B has a paralog, RPL21A, that arose from the whole genome duplication
YPL081W	RPS9A	Protein component of the small (40S) ribosomal subunit; homologous to mammalian ribosomal protein S9 and bacterial S4; RPS9A has a paralog, RPS9B, that arose from the whole genome duplication
YPL089C	RLM1	MADS-box transcription factor; component of the protein kinase C-mediated MAP kinase pathway involved in the maintenance of cell integrity; phosphorylated and activated by the MAP-kinase Slt2p; RLM1 has a paralog, SMP1, that arose from the whole genome duplication

YPL120W	VPS30	Subunit of phosphatidylinositol (PtdIns) 3-kinase complexes I and II; Complex I is essential in autophagy, Complex II is required for vacuolar protein sorting; required for overflow degradation of misfolded proteins when ERAD is saturated; C-terminus has novel globular fold essential for autophagy through the targeting of the PI3-kinase complex I to the pre-autophagosomal structure; ortholog of higher eukaryote gene Beclin 1; human BECN1 can complement yeast null mutant
YPL129W	TAF14	Subunit of TFIID, TFIIF, INO80, SWI/SNF, and NuA3 complexes; involved in RNA polymerase II transcription initiation and in chromatin modification; contains a YEATS domain
YPL143W	RPL33A	Ribosomal 60S subunit protein L33A; N-terminally acetylated; rpl33a null mutant exhibits slow growth while rpl33a rpl33b double null mutant is inviable; homologous to mammalian ribosomal protein L35A, no bacterial homolog; RPL33A has a paralog, RPL33B, that arose from the whole genome duplication
YPL218W	SAR1	ARF family GTPase; component of the COPII vesicle coat; required for transport vesicle formation during ER to Golgi protein transport; lowers membrane rigidity aiding vesicle formation; localizes to ER-mitochondrial contact sites where it enhances membrane curvature, thereby reducing contact size via its N-terminal amphipathic helix; regulates mitochondrial fission and fusion dynamics
YPL283C	YRF1-7	Helicase encoded by the Y' element of subtelomeric regions; highly expressed in the mutants lacking the telomerase component TLC1; potentially phosphorylated by Cdc28p
YPR028W	YOP1	Reticulon-interacting protein; ER integral membrane protein involved in the generation of tubular ER morphology; promotes membrane curvature; forms tubules in vitro; regulates the ER asymmetry-induced inheritance block during ER stress; role in ER-derived peroxisomal biogenesis; interacts with Yip1p to mediate membrane traffic and with Sey1p to maintain ER morphology; facilitates lipid exchange between the ER and mitochondria; forms ER foci upon DNA replication stress
YPR055W	SEC8	Essential 121 kDa subunit of the exocyst complex; the exocyst mediates polarized targeting and tethering of post-Golgi secretory vesicles to active sites of exocytosis at the plasma membrane prior to SNARE-mediated fusion; involved in ER and Golgi inheritance in small buds; relocalizes away from bud neck upon DNA replication stress

YPR172W		Protein of unknown function; predicted to encode a pyridoxal 5'-phosphate synthase based on sequence similarity but purified protein does not possess this activity, nor does it bind flavin mononucleotide (FMN); transcriptionally activated by Yrm1p along with genes involved in multidrug resistance; YPR172W has a paralog, YLR456W, that arose from the whole genome duplication
YPR187W	RPO26	RNA polymerase subunit ABC23; common to RNA polymerases I, II, and III; part of central core; similar to bacterial omega subunit

Appendix E

Table E.1: Performance of the Proteomics Pipeline was assessed by calculating the sensitivity and specificity of detecting FUSP and FSSP. Abbreviations: FUSP (Filtered Unselected SGD Proteins) = Number of proteins in common between 3,613 filtered proteins and 6,270 remaining unselected SGD proteins per group; rep = proteomics experiment biological replicate 1, 2, or 3 (Tyagi and Pedrioli (2015)); FUSP True Positives = number of FUSP below the FDR; FUSP False Positives = number of reversed sequences of FUSP below the FDR; FUSP Sensitivity = sensitivity for detection of FUSP; FUSP Specificity = specificity for detection of FUSP; FSSP (Filtered Selected SGD Proteins) = Number of proteins in common between 3,613 filtered proteins and 330 selected proteins per group; FSSP True Positives = number of FSSP below FDR threshold; FSSP False Positives = number of reversed sequences of FSSP below the FDR; FSSP Sensitivity = sensitivity for detection of FSSP; FSSP Specificity = specificity for detection of FSSP

Group	Rep	Number of FUSP	FUSP True Positives	FUSP False Positives	FUSP Sensitivity	FUSP Specificity	Number of FSSP	FSSP True Positives	FSSP False Positives	FSSP Sensitivity	FSSP Specificity
1	1	3437	3341	6	97.2	99.8	176	2	1	1.1	99.4
	2	3437	3363	14	97.8	99.6	176	4	2	2.3	98.9
	3	3437	3311	3	96.3	99.9	176	3	1	1.7	99.4
2	1	3429	3331	3	97.1	99.9	184	1	4	0.5	97.8
	2	3429	3349	10	97.7	99.7	184	3	5	1.6	97.3

	3	3429	3301	1	96.3	100.0	184	2	5	1.1	97.3
3	1	3438	3341	6	97.2	99.8	175	7	2	4.0	98.9
	2	3438	3362	13	97.8	99.6	175	6	2	3.4	98.9
	3	3438	3310	1	96.3	100.0	175	4	2	2.3	98.9
4	1	3422	3318	4	97.0	99.9	191	3	2	1.6	99.0
	2	3422	3343	9	97.7	99.7	191	1	3	0.5	98.4
	3	3422	3289	2	96.1	99.9	191	1	1	0.5	99.5
5	1	3429	3344	3	97.5	99.9	184	2	3	1.1	98.4
	2	3429	3365	11	98.1	99.7	184	1	2	0.5	98.9
	3	3429	3313	1	96.6	100.0	184	0	1	0.0	99.5
6	1	3420	3327	3	97.3	99.9	193	6	5	3.1	97.4
	2	3420	3341	11	97.7	99.7	193	3	3	1.6	98.4
	3	3420	3291	2	96.2	99.9	193	4	1	2.1	99.5
7	1	3411	3314	2	97.2	99.9	202	4	9	2.0	95.5
	2	3411	3333	12	97.7	99.6	202	4	5	2.0	97.5
	3	3411	3281	0	96.2	100.0	202	3	7	1.5	96.5
8	1	3443	3340	4	97.0	99.9	170	1	1	0.6	99.4
	2	3443	3364	12	97.7	99.7	170	2	1	1.2	99.4
	3	3443	3312	2	96.2	99.9	170	2	1	1.2	99.4
9	1	3431	3336	6	97.2	99.8	182	1	1	0.5	99.5
	2	3431	3353	12	97.7	99.7	182	1	3	0.5	98.4
	3	3431	3303	1	96.3	100.0	182	2	1	1.1	99.5
10	1	3441	3338	6	97.0	99.8	172	2	2	1.2	98.8
	2	3441	3360	10	97.6	99.7	172	3	2	1.7	98.8
	3	3441	3310	2	96.2	99.9	172	2	2	1.2	98.8
11	1	3436	3340	4	97.2	99.9	177	2	0	1.1	100.0
	2	3436	3359	9	97.8	99.7	177	2	0	1.1	100.0

	3	3436	3307	0	96.2	100.0	177	1	0	0.6	100.0
12	1	3433	3334	8	97.1	99.8	180	5	3	2.8	98.3
	2	3433	3354	13	97.7	99.6	180	3	3	1.7	98.3
	3	3433	3304	1	96.2	100.0	180	4	1	2.2	99.4
13	1	3419	3319	5	97.1	99.9	194	4	5	2.1	97.4
	2	3419	3332	10	97.5	99.7	194	5	6	2.6	96.9
	3	3419	3286	1	96.1	100.0	194	3	6	1.5	96.9
14	1	3424	3321	5	97.0	99.9	189	4	4	2.1	97.9
	2	3424	3342	11	97.6	99.7	189	5	4	2.6	97.9
	3	3424	3286	3	96.0	99.9	189	4	3	2.1	98.4
15	1	3444	3340	6	97.0	99.8	169	3	4	1.8	97.6
	2	3444	3363	13	97.6	99.6	169	2	0	1.2	100.0
	3	3444	3314	0	96.2	100.0	169	2	4	1.2	97.6
16	1	3423	3325	5	97.1	99.9	190	2	1	1.1	99.5
	2	3423	3347	13	97.8	99.6	190	2	1	1.1	99.5
	3	3423	3297	2	96.3	99.9	190	2	1	1.1	99.5
17	1	3440	3336	5	97.0	99.9	173	2	1	1.2	99.4
	2	3440	3357	13	97.6	99.6	173	4	2	2.3	98.8
	3	3440	3301	2	96.0	99.9	173	2	2	1.2	98.8
18	1	3441	3345	4	97.2	99.9	172	5	3	2.9	98.3
	2	3441	3362	12	97.7	99.7	172	3	2	1.7	98.8
	3	3441	3310	1	96.2	100.0	172	3	3	1.7	98.3
19	1	3438	3334	6	97.0	99.8	175	3	3	1.7	98.3
	2	3438	3358	14	97.7	99.6	175	5	2	2.9	98.9
	3	3438	3302	1	96.0	100.0	175	4	2	2.3	98.9
20	1	3448	3353	4	97.2	99.9	165	2	4	1.2	97.6
	2	3448	3374	12	97.9	99.7	165	2	4	1.2	97.6

	3	3448	3324	1	96.4	100.0	165	1	4	0.6	97.6
Average		3432.4	3330.2	5.9	97.0	99.8	180.7	2.9	2.6	1.6	98.6